EIRÍKUR RÖGNVALDSSON
AUÐUR ÞÓRUNN RÖGNVALDSDÓTTIR
KRISTÍN BJARNADÓTTIR
SIGRÚN HELGADÓTTIR

Vélræn málfræðigreining með námfúsum markara

1. Mörkun og markarar

1.1 Hvað er mörkun?

Með *mörkun* (e. tagging) er átt við það að merkja eindir í samfelldum texta á kerfisbundinn hátt. Þessar merkingar geta verið af ýmsum toga. Þannig er t.d. hægt að hugsa sér að öll mannanöfn séu merkt á ákveðinn hátt, öll staðanöfn á annan hátt, öll erlend orð í textanum séu sérmerkt, o.s.frv. Það sem hér skiptir máli er mörkun málfræðilegra upplýsinga. Grundvallaratriðið er þar *orðflokksmörkun* (e. PoS tagging), þar sem orðflokksmerki er hengt á hvert orð, t.d. *Gamla*<lo> *konan*<no> *mætti*<so> *þessum*<fn> *tveim*<to> *drengjum*<no> *í*<fs> *morgun*<no>. Síðan er hægt að ganga lengra og bæta inn hvers kyns málfræðilegum upplýsingum, s.s. um kyn, tölu, fall, persónu, tíð, stig o.s.frv. Einnig má marka orð með upplýsingum um setningafræðileg hlutverk, s.s. frumlag, andlag, umsögn o.þ.h.

Málfræðileg greining og mörkun er nauðsynleg í margvíslegum tungutæknitólum. Meginhluti málfarsleiðréttinga er t.d. óhugsandi án slíkrar greiningar. Aðeins lítill hluti málfarsvillna felst í því að notaðar séu orðmyndir sem ekki eiga að koma fyrir í málinu (t.d. *föður**s*** í stað *föður, k**ey**ptu* í stað *k**au**ptu*). Langflestar villur felast í því að nota leyfilegar orðmyndir á óleyfilegum stöðum í setningu. Villur eins og *Ég hitti syst**ir** þína* (í stað *syst**ur***), *vegna þeirrar tilhneiging**u*** (í stað *tilhneiging**ar***), *fjöldi manna kom**u*** (í stað *kom*), ***mér** langar* (í stað ***mig** langar*) o.s.frv. er ekki hægt að finna án málfræðilegrar greiningar, því að *systir*, *tilhneigingu*, *mér* og *komu* eru allt fullkomlega leyfilegar íslenskar orðmyndir – bara ekki á þessum stöðum í setningu.

Ýmsar algengar stafsetningarvillur eru líka þess eðlis að þær finnast ekki nema með málfræðilegri greiningu. Mörg orð í málinu eru t.d. ýmist skrifuð með einu eða tveimur *n*-um eftir setningafræðilegri stöðu; *morgunn*/*morgun*, *Kristinn*/*Kristin*, *farinn*/*farin* o.s.frv. Hér eru báðar myndirnar leyfilegar, og villuleitarforrit sem eingöngu skoðar hverja orðmynd fyrir sig finnur því ekki villurnar í *það er kominn morgu**n**, ég hitti Kristi**nn**, hann er fari**n***.

Vélrænar þýðingar byggjast einnig á málfræðilegri greiningu. Án slíkrar greiningar getur vélræn þýðing ekki orðið annað en einföld uppfletting í orðasafni, þar sem orð úr einu máli er sett í stað orðs í öðru máli og ekkert hirt um reglur um orðaröð, beygingar og annað slíkt. Þá koma upp alþekkt dæmi eins og *hot spring river this book* (fyrir *hver á þessa bók*). Þar fyrir utan er málfræðimörkun vitanlega mjög gagnleg í orðabókagerð og á ýmsum öðrum sviðum sem óþarfi er að rekja.

1.2 Hvernig fer mörkun fram?

Til að marka texta þarf sérstakt forrit, *markara* (e. tagger). Málfræðileg mörkun fer venjulega fram í tveimur þrepum. Í fyrra þrepinu er orðum í textanum flett upp í orða-safni með beygingarlegum upplýsingum, og þær upplýsingar síðan færðar inn í textann. Þannig eiga t.d. að fást upplýsingar um að *í* sé forsetning, *hesturinn* sé nafnorð í karlkyni, eintölu, nefnifalli, með greini, og *fóruð* sé sögn í annarri persónu, fleirtölu, þátíð, framsöguhætti, germynd. Þessar upplýsingar eru síðan færðar inn í textann.

Slík uppfletting dugir hins vegar ekki til að greina öll orð í samfelldum texta á ótvíræðan hátt. Það kemur t.d. í ljós við uppflettinguna að þótt *í* sé einrætt orð er *á* ekki bara forsetning, heldur líka sögnin *eiga* í 1. og 3. persónu, eintölu, nútíð, framsöguhætti, germynd; kvenkynsnafnorðið *á* í eintölu, nefnifalli, þolfalli og þágufalli; kvenkynsnafnorðið *ær* í eintölu, þolfalli og þágufalli; og fleira mætti nefna. Þótt greiningin á *hestur* sé ótvíræð getur *hesta* verið bæði þolfall og eignarfall fleirtölu. Þótt *fóruð* sé einrætt er *fórum* tvírætt; getur ekki einungis verið fyrsta persóna, fleirtala, þátíð, framsöguháttur, germynd af *fara*, heldur líka þágufall fleirtölu af *fórur* (sem reyndar kemur tæpast fyrir í eintölu).

Til að leysa úr slíkri tví- og margræðni þarf annað þrep í vinnslunni. Í því lagi er önnur eða ein greiningin valin, en hinni eða hinum hafnað. Forrit sem framkvæma slíkt val vinna á ýmsa vegu, en í grundvallaratriðum má segja að þau skiptist í tvo flokka; *reglumarkara* (e. rule-based taggers) og *tölfræðimarkara* (e. statistical/ stochastic taggers). Þriðja tegundin er svo *námfúsir markarar* (e. transformation-based taggers), sem nýta sér bæði reglur og tölfræðilegar upplýsingar, og þetta tvennt getur spilað saman á margvíslegan hátt (sjá Jurafsky & Martin 2000:300-312).

Reglumarkarar nota reglur um gerð setninga og setningarliða til að marka orðin. Þeir búa t.d. yfir upplýsingum um það að forsetning kemur sjaldan næst á undan sögn, og þess vegna er ólíklegt að orðið *fórum* sé sögn í sambandinu *í fórum mínum*, Þótt svo gæti verið ef litið er á orðið eitt og sér. Reglumarkari ætti líka að búa yfir upplýsingum um það að þegar eignarfornafn stendur næst á eftir nafnorði sambeygjast orðin venjulega; þ.e., standa í sama kyni, tölu og falli. Í sambandinu *hesta þinna* er *þinna* ótvírætt eignarfall, og þær upplýsingar eiga að nægja til að úrskurða að *hesta* sé líka eignarfall, en ekki þolfall eins og það gæti einnig verið ef litið er á orðið eitt og sér.

Tölfræðimarkarar byggjast á upplýsingum um tíðni einstakra beygingarmynda til að velja líklegustu greininguna. Slíkur markari myndi greina *á* rétt í setningunni *Ég er á leiðinni*, vegna þess að *á* er mun oftar forsetning en nokkuð annað. Hins vegar yrði *á* ranglega greint í setningunni *Ég á þetta*; þar veldi tölfræðimarkarinn forsetningu eins og áður. Sömuleiðis yrði *fórum* trúlega greint ranglega sem sögn í sambandinu *í fórum mínum*, því að þessi orðmynd er mun algengari sem sagnmynd en sem nafnorðsmynd.

Báðar þessar tegundir hafa kosti og galla. Tölfræðimarkarar hafa þann kost að það er tiltölulega fljótlegt að koma þeim upp. Anders Nøklestad frá Tekstlaboratoriet í Osló nefndi það í fyrirlestri sínum á ráðstefnunni *Sambúð tungu og tækni* 13. nóvember sl. að hann hefði gert tölfræðimarkara fyrir norsku á þremur mánuðum. Það fóru hins vegar fjögur ársverk í norska reglumarkarann (sjá einnig Nøklestad 1998). Á hinn bóginn er hægt að ná betri niðurstöðum, þ.e. færri röngum greiningum, með reglumörkurum en tölfræðimörkurum; og reglumarkarar ráða betur við margbrotna greiningu en tölfræðimarkarar.

Það er þó sameiginlegt flestum mörkurum að þeir þurfa á að halda sérstöku *þjálfunarsafni* (e. training corpus). Það er texti sem hefur verið greindur handvirkt

eftir sama kerfi og vélræna greiningin á að nota. Við gerð reglumarkara nýtist þetta safn til að átta sig á þeim mynstrum í textanum sem hægt er að setja fram í regluformi; við gerð tölfræðimarkara nýtist safnið til að afla upplýsinga um tíðni einstakra orðmynda, og tíðni mismunandi greininga á sömu orðmynd. Nauðsynlegt er að þjálfunarsafnið sé stórt og samsett úr fjölbreyttum textum til að markarinn sem byggist á því skili sem réttustum niðurstöðum.

1.3 Markari Brills

Ein þekktasta útfærslan á mörkurum er kennd við Eric Brill sem nú er sérfræðingur hjá Microsoft, og yfirleitt nefnd *Brill's tagger*, *Brill type tagger* eða eitthvað í þá átt (sjá Brill 1995). Slíkur markari byggist á aðferð sem nefnist *transformation based learning*. Með því er átt við það að markarinn er keyrður á þjálfunarsafn þar sem hvert orð hefur tvo (eða hugsanlega fleiri) greiningarstrengi. Hjá einræðum orðmyndum eru báðir (eða allir) strengirnir samhljóða; en ef orðmynd er fleirræð fær hún fleiri greiningarstrengi. Verkefni markarans er svo að finna aðferðir til að velja rétta strenginn. Eftirfarandi dæmi er tekið úr *The Wall Street Journal Corpus*:

```
(1)   wd(7799,a).
      tag(7799,'DT').
      tag('DT','DT',7799).

      wd(7800,good).
      tag(7800,'JJ').
      tag('JJ','JJ',7800).

      wd(7801,buy).
      tag(7801,'VB').
      tag('VB','NN',7801).
```

Hér táknar *wd* orð og *tag* greininguna; *7799*, *7800* og *7801* eru svo bara hlaupandi númer orðanna í textanum. Af þessu má ráða að *a* er einrætt orð og getur aðeins verið greinir (*DT*); *good* er einnig einrætt og getur aðeins verið lýsingarorð (*JJ*); en *buy* getur aftur á móti hvort heldur er verið sögn (*VB*) eða nafnorð (*NN*). Hér hefur greiningin verið yfirfarin handvirkt og séð til þess að ranga greiningin kemur ævinlega á undan þeirri réttu.

Markarinn les nú skrána og stansar við orð sem fá tvo mismunandi greiningarstrengi, eins og *buy* hér að framan. Þar fær hann þær upplýsingar að rétt greining orðmyndarinnar sé nafnorð, en í einangrun gæti hún eins verið sögn. Þá leitar hann að einhverju í umhverfinu sem gefur honum vísbendingar um að hér sé um nafnorð að ræða, en ekki sögn. Í þeirri leit hefur hann hliðsjón af nokkrum *sniðmátum* (e. templates), sem leiðbeina um það hvað í umhverfinu gæti skipt máli. Eitt slíkt sniðmát gæti verið:

```
(2)   tag:A>B <- tag:C@[-1].
```

Þetta lesist: Breytið greiningarstreng *A* í greiningarstreng *B* ef undanfarandi orð hefur greiningarstreng *C*. Setjum nú svo að annars staðar í textanum komi fyrir orðarunan *a bad taste*, og *taste* fái einnig tvöfalda greiningu, sem sögn og nafnorð. Þá getur markarinn ályktað sem svo að hér sé fundin regla um val milli sagnar og nafnorðs í ákveðnu umhverfi, og búið til eftirfarandi reglu út frá sniðmátinu hér að framan:

```
(3)   tag:VB>NN <- tag:JJ@[-1].
```

Þessi regla segir: Breytið greiningunni *sögn* í greininguna *nafnorð* ef orðið á undan er *lýsingarorð*.

Ef markarinn er keyrður á sæmilega stórt þjálfunarsafn kemur hann sér upp talsverðum fjölda reglna af þessu tagi. Það er háð ýmsum breytum hversu margar reglurnar verða. Það fer m.a. eftir því hversu oft tiltekið samband kemur fyrir. Dugir að það komi fyrir tvisvar í textanum, eins og í dæminu af *a good buy* og *a bad taste* hér að framan, eða þarf það að koma oftar fyrir til að vera talið regla fremur en tilviljun – og þá hversu oft? Reglufjöldinn fer einnig eftir fjölda og gerð sniðmáta. Hér að framan var aðeins sýnt eitt sniðmát, þar sem reglan miðast við greiningu undanfarandi orðs; en einnig má hugsa sér sniðmát á við þessi:

```
(4)   tag:A>B <- tag:C@[-1] & tag:D@[1].
      tag:A>B <- wd:C@[1].
```

Fyrra sniðmátið segir: Breytið greiningarstreng *A* í *B* ef undanfarandi greiningarstrengur er *C* og eftirfarandi greiningarstrengur er *D*. Seinna sniðmátið segir: Breytið *A* í *B* ef orðið á eftir er *C*.

Það reglusafn sem til verður við þetta er síðan keyrt á sérstakt *prófunarsafn* (e. test corpus) sem er texti með réttri greiningu allra orða. Þá er hægt að meta hversu fullkomið reglusafnið, út frá því hversu oft það skilar sömu greiningu og orðin hafa í prófunarsafninu. Nauðsynlegt er að skoða hvaða villur markarinn gerir og reyna síðan að endurbæta hann með því t.d. að fjölga og breyta sniðmátum. Stærð þjálfunarsafnsins skiptir einnig miklu máli; eftir því sem það er stærra má búast við betri niðurstöðum.

Þegar búið er að koma upp eins fullkomnu reglusafni og hægt er þarf að skrifa tvö forrit, áður en farið er út í að marka áður ómarkaða texta. Í fyrsta lagi er það svonefndur *giskari* (e. unknown word guesser). Hann er nauðsynlegur vegna þess að ekki er hægt að gera ráð fyrir því að öll orð í textanum sem verið er að marka finnist í orðasafninu sem unnið er með. Þessi giskari reynir þá að greina óþekkt orð út frá endingum, viðskeytum og öðrum atriðum sem geta verið til leiðbeiningar um greiningu. Í öðru lagi þarf að skrifa markarann sjálfan. Hann flettir upp í orðasafninu, skrifar mögulegar greiningar hverrar orðmyndar inn í textann, og velur réttu greininguna úr þeim hópi í samræmi við reglusafnið.

2. Gerð íslensks markara

2.1 Hráefni og undirbúningur

Við höfum undanfarið unnið að undirbúningi þess að þróa markara sem geti greint íslenskan texta málfræðilega með a.m.k. 95% nákvæmni. Ætlunin er að nota sérstaka útfærslu af markara Brills sem Torbjörn Lager, kennari í tungutækni við Gautaborgarháskóla, hefur skrifað og kallar *μ-tbl* (sjá Lager 1999). Við höfum unnið nokkuð með þennan markara og fengið reynslu af því að endurbæta hann. Einnig höfum við gert tilraun með að nota hann á íslenskt efni. Að fenginni þeirri reynslu teljum við ótvírætt að þessi markari geti nýst vel við mörkun á íslenskum textum. En forsenda fyrir því að hann nýtist er sú að til er afbragðsgott þjálfunarsafn. Það eru grunnskrárnar úr vinnslu *Íslenskrar orðtíðnibókar*, sem Orðabók Háskólans gaf út 1991. Ritstjóri bókarinnar var Jörgen Pind, en Stefán Briem sá um vélræna málfræðigreiningu og Friðrik Magnússon um handvirka greiningu.

Form skránna er sýnt í (5). Fremst er greiningarstrengur sem inniheldur upplýsingar um orðflokk og öll beygingarleg atriði. Strengurinn *n k e n g* fyrir framan *hvolpurinn* merkir þannig ***n****afnorð,* ***k****arlkyn,* ***e****intala,* ***n****efnifall,* ***g****reinir*.

```
(5)   f p k e n     hann        hann
      s f g 3 e þ o átti        eiga
      n h e o       afmæli      afmæli
      a o           í           í
      n k e o       dag         dagur
      c             og          og
      n k e n g     hvolpurinn  hvolpur
      n k e n - m   Vaskur      Vaskur
      s f g 3 e þ   var         vera
      n v e n       afmælisgjöf afmælisgjöf
```

Þessi greining var að nokkru leyti unnin vélrænt, en síðan var farið vandlega yfir hana alla í höndunum og það á að vera hægt að treysta því að hún sé rétt. Þetta hráefni er alls 500 þúsund orð (5000 orða bútar úr 100 textum, sem skiptast á fimm mismunandi efnisflokka). Hér er því um að ræða mjög stórt og sérlega verðmætt þjálfunarsafn (til samanburðar má nefna að þjálfunarsafnið í hinu norska *taggerprosjekt* var um 100 þúsund orð, og textarnir í því ekki sérstaklega valdir).

Greiningin í *Íslenskri orðtíðnibók* er mjög nákvæm; það er notuð stór markaskrá (e. *tagset*). T.d. er fallstjórn forsetninga og sagna greind sérstaklega; upplýsingar um fallstjórn sagna birtast þó ekki í prentuðu bókinni. Alls kemur 621 mismunandi greiningarstrengur fyrir í bókinni.

Byrjað á að taka öll orðin í grunnskrám *Orðtíðnibókarinnar* og raða þeim í stafrófsröð. Mörg þeirra fá þá fleiri en einn greiningarstreng. Þá er algengasti strengurinn tekinn og keyrður sem aukastrengur inn í markaða textann, á undan rétta greiningarstrengnum. Í mjög mörgum tilvikum verður aukastrengurinn sá sami og hinn rétti greiningarstrengur sem orðið hefur fyrir. En það er auðvitað ekki alltaf sem algengasta greiningin á við, og í þeim tilvikum verða greiningarstrengirnir tveir mismunandi. Athugið þó að alltaf er hægt að sjá hvor greiningin er rétt, vegna þess að upphaflegi (rétti) strengurinn er aftast.

Það kann að virðast undarlegt að byrja á því að bæta röngum greiningum inn í skrá sem er rétt greind. En þetta er nauðsynlegt til að markarinn geti lært reglur sem endurskoða greiningu út frá umhverfi. Þegar hrár texti er markaður frá grunni þarf að byrja á að keyra hann saman við orðasafn með beygingarlegum upplýsingum, eins og áður var nefnt. Þegar um tvíræða orðmynd er að ræða fær hún þá í upphafi tvo greiningarstrengi. Með þeirri aðferð sem lýst er hér að framan lærir markarinn hvernig hann á að bregðast við slíkum aðstæðum; hvernig hann getur farið að því að taka annan strenginn fram yfir hinn.

2.2 Tilraun með mörkun íslensks texta

Við höfum nú tekið sýnishorn af *Orðtíðnibókinni*, tæp 60 þúsund orð, og meðhöndlað þau eins og lýst er hér að framan. Þjálfunarsafnið var tæplega 48 þúsund orð, en prófunarsafnið 11923 orð. Þar af höfðu 2445, eða 79,5%, aðeins einn greiningar-streng, en hjá afgangnum, 9478 orðum, var viðbótarstrengurinn sem keyrður hafði verið inn í skrána annar en hinn rétti greiningarstrengur. Eitt dæmi um það er sýnt hér:

```
(6)   wd(38,'til').
      tag(38,'ae').
      tag('ae','ae',38).

      wd(39,'enda').
      tag(39,'c').
      tag('c','nkee',39).
```

Hér hefur orðmyndin *enda* fengið viðbótarstrenginn *c*, þ.e. samtenging, vegna þess að það er algengasta greining þeirrar orðmyndar í *Orðtíðnibókinni* (189 dæmi). Samhengið sýnir hins vegar ljóslega að rétta greiningin er hér *nkee*, þ.e. no., kk., et., ef. (en sá greiningarstrengur á aðeins 7 sinnum við þessa orðmynd í *Orðtíðnibókinni*).

```
(7)   tag:A>B <- tag:C@[-1].
      tag:A>B <- tag:C@[1].
      tag:A>B <- tag:C@[-1,-2].
      tag:A>B <- tag:C@[-1,-2,-3].
      tag:A>B <- tag:C@[-1] & tag:D@[1].
      tag:A>B <- tag:C@[-1] & tag:D@[-2].
      tag:A>B <- tag:C@[-1] & tag:D@[-2] & tag:E@[-3].
      tag:A>B <- tag:C@[1,2].
      tag:A>B <- tag:C@[-1] & tag:D@[1,2].
      tag:A>B <- wd:C@[0].
      tag:A>B <- wd:C@[1].
      tag:A>B <- wd:C@[-1].
      tag:A>B <- wd:C@[0] & wd:D@[-1].
      tag:A>B <- wd:C@[0] & tag:D@[-1].
      tag:A>B <- wd:C@[0] & tag:D@[1].
      tag:A>B <- wd:C@[-1,-2].
      tag:A>B <- wd:C@[0] & wd:D@[-1] & wd:E@[-2].
```

Í (7) sjáum við svo þau sniðmát sem við notuðum. Þau eru alls 17; níu af þeim vísa eingöngu til marka, þ.e. málfræðilegrar greiningar orðanna í kring, sex vísa eingöngu til orða, en tvö vísa bæði til marka og orða. Mínus á undan tölu táknar að vísað er til undanfarandi orðs eða marks, en sé enginn mínus er vísað til eftirfarandi orðs eða marks. Komma á milli talna táknar ‘annaðhvort’, en & táknar ‘hvorttveggja’.
Eftir að aukagreiningarstreng hafði verið bætt inn í textann voru 79,5% orðanna í prófunarsafninu með ótvíræða greiningu, þannig að algengasta greining þeirra orðmynda, sem bætt var inn, var hin sama og rétta greiningin, sem fyrir var. Eftir að *μ-tbl* forritið hafði verið keyrt þrisvar á þjálfunarsafnið og lært alls 609 reglur var það keyrt á prófunarsafnið. Þá tókst því að fækka röngum greiningum niður í 1026, þannig að 91,5% greiningarstrengja voru réttir.

Síðan gerðum við tilraunir með að minnka markaskrána, þ.e. einfalda greininguna nokkuð, og tókum út upplýsingar um fallstjórn sagna og forsetninga. Ástæðan fyrir því er sú að þetta eru atriði sem vanalega eru ekki tiltekin í málfræðilegri greiningu, og þau verða væntanlega ekki heldur fyrir hendi í þeirri beygingarlýsingu sem við vonumst til að geta notað þegar þar að kemur. Við þetta fækkaði tvíræðum greiningum í prófunarsafninu talsvert, þannig að lagt var upp með 89% ótvíræða greiningu þar. Forritið var svo keyrt þrisvar á þjálfunarsafnið og lærði alls 339 reglur. Þær reglur voru svo keyrðar á prófunarsafnið og fækkuðu röngum greiningum í 616. Það þýðir að 95% greiningarstrengja eru orðnir réttir.

Þetta er að okkar mati ótrúlega góður árangur af fyrstu tilraun. Þó verður að hafa í huga að eftir því sem hlutfall réttra greiningarstrengja er orðið hærra verður erfiðara að bæta niðurstöðuna. Það er enn langur vegur upp í 98% rétta greiningu, en hærra verður tæplega komist. Ástæðan fyrir því er sú að eftir það fer málfræðinga að greina á. Er *sem* t.d. tilvísunarfornafn eða tilvísunartenging? Er *gær* atviksorð eða nafnorð? Er alltaf hægt að greina hvort sögn er í framsöguhætti eða viðtengingarhætti? O.s.frv.

Við höfum líka séð það á tilraunum sem við höfum gert að nákvæm greining hefur bæði kosti og galla. Það er t.d. oft erfitt að greina fallstjórn sagna og forsetninga, og margar villur í fyrstu greiningu okkar stöfuðu af því að fallstjórnin var rangt

greind. En ef það tekst að greina fallstjórnina, þá kemur sú greining að miklum notum við greiningu á falli orðanna sem þessar sagnir og forsetningar stýra.

2.3 Reglur og villur

Lítum nú aðeins á þær reglur sem markarinn dró út úr þjálfunarsafninu. Í (8) sjáum við nokkur dæmi um reglur sem hann lærði í seinni tilrauninni.

```
(8)   tag:sfg3eþ>sfg1eþ <- tag:fp1en@[-1,-2] o
      tag:cn>c <- tag:svg3en@[1,2] o
      tag:cn>c <- tag:svg3eþ@[1,2] o
      tag:af>fp1fn <- wd:við@[0] & tag:sfg1fn@[1] o
      tag:sfg3en>sfg1en <- tag:fp1en@[-1,-2] o
      tag:cn>c <- tag:sfg3eþ@[1,2] o
      tag:af>fp1fn <- wd:við@[0] & tag:sfg1fþ@[1] o
      tag:sfg3eþ>sfg1eþ <- tag:fp1en@[1] o
      tag:svg3eþ>svg1eþ <- tag:fp1en@[-1] o
      tag:fpken>fpkeo <- tag:af@[-1] o
      tag:cn>c <- tag:sfg3en@[1,2] o
      tag:sfg3en>sfg2en <- tag:fp2en@[-1,-2] o
      tag:ssg>sþghen <- wd:var@[-1,-2] o
      tag:foheþ>lheþsf <- wd:einu@[0] & wd:í@[-1] o
      tag:fahen>faheo <- tag:af@[-1] o
      tag:af>fp1fn <- wd:við@[0] & tag:sfg1fn@[-1] o
```

Fyrsta reglan segir: Breytið greiningunni *sögn, framsöguháttur, germynd, þriðja persóna, eintala, þátíð* í *sögn, framsöguháttur, germynd, fyrsta persóna, eintala, þátíð* ef greining næsta eða þarnæsta orðs á undan er *fornafn 1. persónu, eintala, nefnifall* – þ.e. *ég*. 7. reglan segir: Breytið greiningunni *atviksorð/fornafn* í *fornafn 1. persónu, fleirtölu, nefnifall* ef orðið er *við* og greining eftirfarandi orðs er *sögn, framsöguháttur, germynd, fyrsta persóna, fleirtala, þátíð*. Síðasta reglan er alveg eins, nema hún vísar til nútíðarmynda sagna í stað þátíðar. Svo sjáum við tvær reglur sem breyta greiningu sagna úr 3. persónu í fyrstu ef fyrstu persónu fornafn fer næst á undan eða eftir; og svo mætti halda áfram.

Lítum aðeins á nokkur dæmi um algengustu villurnar í greiningunni. Þetta eru allt dæmi úr seinni greiningunni, þegar búið var að taka fallstjórnina út. Í (9) sést að 10 dæmi voru um það að nafnorð í þágufalli (*þ*) væri greint sem þolfall (*o*).

(9) **10 dæmi um *nveo* sem ættu að vera *nveþ*:**

327:	að þeir séu sleipir í sögu . það er eiginlega ekki
489:	að fara í andaglas í tölvu . hvað þessir strákar þykjast
1372:	ofan í sig , þessari túpu sem átti að liggja slétt
3130:	sem hún fylgdist með hverri hreyfingu hetju sinnar . Alli gaf
5883:	einhverjum ástæðum var ömmu í nöp við þessa iðju . Jóra
6586:	týndist Stóri-Jón reyndi eftir bestu getu að segja frá ferðalaginu niður
10351:	súkkulaði og kaffi inni í stofu . meira að segja Guðmundur
4101:	ætlaði að koma við í sjoppu og . kaupa eitthvert sælgæti
5469:	. einar dyr voru á framhlið hans en þær voru sjaldan
5049:	var farið að svima af eftirvæntingu . hún rétti út höndina

Í öllum þessum dæmum nema einu fer forsetning á undan nafnorðinu, þó ekki alltaf næst á undan – ég hef auðkennt forsetningarnar hér til glöggvunar. Vandinn er sá að þessar forsetningar stjórna ýmist þolfalli eða þágufalli. Auðvitað er hægt að ráða það

af samhenginu hvort fallið er rétt í hverju tilviki, en mér finnst ekki augljóst að hægt sé að semja reglur sem segi til um greininguna. Þær reglur yrðu oft ansi flóknar, og eins líklegt að þær gætu af sér fleiri rangar greiningar en réttar.

(10) **9 dæmi um *fphen* sem ættu að vera *fpheo*:**

6662: átti ég að vita , það muldraði . lögregluþjónninn má ég
8193: undir fótum sér , sá það , heyrði það , fékk
8199: , heyrði það , fékk það í fangið þegar hann datt
9864: þrjóskur . flugmaðurinn sagði mér það . þá lýgur hann því
10563: skotinn í henni , sérðu það ekki ? hvíslaði Kata spekingslega
10600: í bók að maður sjái það á augunum í fólki .
10873: jólafríið kemur . kennarinn segir það . ætlar hann að kenna
11426: já . amma mín kallaði það að krossa sig , en
11870: skapað líf . Jesús getur það . af því að hann

Hér eru svo dæmi um að fornafn 3. persónu í hvorugkyni, *það*, sé greint sem nefnifall þar sem það er í raun þolfall. Í öllum dæmunum fer sögn á undan *það*, og því gæti manni virst það vera tiltölulega einfalt að setja fram reglu sem segði að þegar *það* kæmi á eftir sögn bæri að greina það sem þolfall. Gallinn er hins vegar sá að reglurnar geta ekki vísað til orðflokka, heldur vísa þær til greiningarstrengsins í heild. Það þýðir að eina reglu þarf til að segja til um fallið á *það* á eftir sögn í 1. persónu eintölu framsöguhætti nútíð germynd, aðra reglu fyrir sögn í 2. persónu eintölu framsöguhætti nútíð germynd, o.s.frv. Þess vegna eru dæmin um hvern greiningarstreng í þjálfunarsafninu ekki nægilega mörg til þess að markarinn læri neina reglu. Ef hægt væri að vísa til hluta strengsins, t.d. bara fyrsta stafsins í honum sem táknar orðflokk, myndi þetta gerbreytast.

3. Lokaorð

Þótt enn standi vissulega talsvert eftir af villum teljum við að góðir möguleikar séu á að ná mun betri árangri í greiningunni, enda eigum við enn mörg tromp uppi í erminni. Meðal þess sem unnt er að gera til að bæta árangurinn er að:

1. Stækka þjálfunarsafnið. Eins og áður segir nýtum við nú aðeins um 1/10 af *Orðtíðnibókinni* sem þjálfunarsafn, en gætum nýtt allt að 9/10 (og afganginn þá sem prófunarsafn). Því stærra sem þjálfunarsafnið er, þeim mun fleiri og betri reglur verða til.
2. Fjölga sniðmátum og endurbæta þau. Eins og áður kom fram ákvarða sniðmátin form reglnanna. Hér er hugsanlegt að fram komi munur á íslensku og ýmsum öðrum málum. E.t.v. hafa beygingar í íslensku þau áhrif að þar þurfi að skoða stærra umhverfi (t.d. þrjú orð á undan og eftir).
3. Einfalda greininguna. Greiningin í *Orðtíðnibókinni* er mun nákvæmari en venja er við vélræna mörkun; greiningarstrengir eru alls 621. Með því að einfalda greininguna er hægt að ná betri árangri; vega þarf og meta hversu mikilvæg einstök greiningaratriði eru.
4. Lagfæra reglurnar eftir á. Forritið skilar út skrá um þær villur sem standa eftir í prófunarsafninu, eftir að reglusafnið hefur verið keyrt á það. Með því að skoða þessar villur má oft sjá regluleika sem forritið hefur ekki bundið í reglur af einhverjum ástæðum, og búa slíkar reglur til handvirkt.

Þegar búið er að ná eins góðum niðurstöðum úr markaranum og mögulegt er, með því að þjálfa hann á grunnskrám *Orðtíðnibókarinnar*, má byrja á að nota hann á ómarkaða texta. Við vonumst til að komast upp í 98% rétta greiningu áður en yfir lýkur, þótt við gerum okkur ljóst að það geti orðið erfitt. En niðurstaða í því máli fæst vonandi síðar á þessu ári.

Heimildir

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* 21: 543-566.

Jurafsky, Daniel, & James H. Martin. 2000. *Speech and Language Processing.* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey.

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

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

Nøklestad, Anders. 1998: Statistisk disambiguerende tagging av norsk. Jan Terje Faarlund, Britt Mæhlum & Torbjørn Nordgård (ritstj.): *MONS 7. Utvalde artiklar frå det 7. møtet om norsk språk i Trondheim 1997.* Novus Forlag, Osló.