Þórdís Úlfarsdóttir

Málfræðileg mörkun orðasambanda

1 Inngangur

Hér verður lýst tilraun til að marka orðasambönd eins og þau sem eru í orðasambandaskrá Orðabókar Háskólans.[1] Að *marka* er það kallað þegar tölvutækur texti er greindur málfræðilega á vélrænan hátt og greiningarstrengur látinn fylgja hverju orði. Forritið sem notað er til þess nefnist *markari* og verknaðurinn heitir *mörkun.* Orðasamböndin sem um ræðir eru af margvíslegu tagi, orðasamband getur til dæmis verið lýsingarorð sem stendur með nafnorði: *feitt embætti,* sögn með sagnfyllingu: *vera glaður og sviplléttur,* eða orðtak: *aka seglum eftir vindi.* Þessi þrjú orðasambönd eru meðal þeirra sem prófað var að marka.

Málfræðileg greining orðasambanda er í því fólgin að tekin eru valin sambönd og þau undirbúin fyrir vélræna greiningu, m.a. með því að setja aðgreinandi tákn á milli þeirra. Síðan er markarinn látinn greina málfræðilega einstök orð í samböndunum. Orðasambönd hafa ekki áður verið greind á þennan hátt í íslensku svo vitað sé.

Byrjað var að marka orðasamböndin í mars 2005 og enda þótt ekki væri farið af stað með ákveðið markmið í huga varð fljótlega ljóst að mörkunin mundi gefa forvitnilegar niðurstöður. Skömmu áður hafði Sigrún Helgadóttir skilað lokaskýrslu um verkefnið **Markarann** sem var nýtt tungutæknitól fyrir íslensku, en hún hafði tekið að sér tæknilega umsjón forritsins og þjálfun þess fyrir íslenska texta hjá Orðabók

[1] Jón Hilmar Jónsson hefur haft umsjón með orðasambandaskránni og hugmyndin að því að marka orðasamböndin er frá honum komin.

Orð og tunga 8 (2006), 117–144.

Háskólans. Með þetta verkfæri í höndum opnuðust nýir möguleikar til málrannsókna því nú var hægt að greina ýmsar tegundir texta í málfræðilegar einingar á skjótan hátt með aðstoð tölvu. Áhugavert var að sjá hvað kæmi út úr mörkun orðasambanda og athuga hvort og hvernig hægt væri að hagnýta þær niðurstöður. Einn ávinningur af þessari tilraun var að mögulegt var að raða orðasamböndunum á nýjan hátt, ekki eftir stafrófsröð heldur eftir greiningarstrengjunum, og komu þá í ljós ýmis forvitnileg setningarleg mynstur sem orðasamböndin mynduðu.

Efnisskipan greinarinnar er þannig að í 2. kafla er fjallað um orðasamböndin og í 3. kafla um markarann og mörkunina. Í 4. kafla er ferli mörkunar lýst og í 5. kafla segir frá nákvæmni mörkunar og helstu takmörkunum hennar. Í 6. kafla eru listar með mörkuðum orðasamböndum og þar er jafnframt greint frá niðurstöðum mörkunarinnar. Í 7. kafla er lýst hvernig hægt er að raða orðasamböndum eftir málfræðigreiningunni, og 8. kafli fjallar um það þegar markaðar eru nokkrar línur af skýringartexta orðabókar. Niðurstöður eru síðan dregnar fram í lokaorðum í 9. kafla.

2 Orðasamböndin

Orðasambönd í orðasambandaskrá Orðabókar Háskólans eru nú um 135 þúsund talsins. Þau eiga upptök sín í Ritmálsskrá Orðabókar Háskólans og eru úrvinnsla úr dæmasafni hennar. Jón Hilmar Jónsson hefur byggt upp orðasambandaskrána sem er aðgengileg á vefnum. Hægt er að fletta upp í skránni á www.lexis.hi.is/osamb/osamb.pl og þar er einnig að finna upplýsingar um hana.

Orðasamböndin hafa að geyma mikinn og fjölbreytilegan orðaforða, í þeim 135 þúsund orðasamböndum sem nú eru tiltæk koma fyrir rúmlega 57.500 mismunandi lykilorð (þ.e. orð sem mynda merkingarkjarna sambandsins). Orðasambandaskránni er ætlað að veita yfirsýn um föst orðasambönd og tilbrigði þeirra og endurspegla dæmigerða notkun orðanna auk þess sem hún veitir innsýn í sögu og samhengi íslensks orðaforða á síðari öldum. Í skránni er að finna orðasambönd af margvíslegu tagi, allt frá föstum samböndum í óbreytanlegri mynd *(gera sér glaðan dag)* til lauslegra orðastæðna *(aka á hundasleða)*.

Á dæmunum fyrir neðan má sjá að orðasamböndin eru ekki "setningar" þar sem í aðeins sumum þeirra er sögn í persónuhætti, og oft

er engin sögn í sambandinu. Til að lýsa nánar því efni sem orðasambandaskráin hefur að geyma eru sýnd hér nokkur sambönd sem tengjast ákveðnum lykilorðum, fyrst koma dæmigerð sambönd með sögn en svo með lýsingarorði. Þessi orðasambönd voru meðal þeirra sem mörkunartilraunin náði yfir.

16 orðasambönd með sögninni aka:

aka angri á bug
aka á hundasleða
aka á tún
aka á völl
aka á þvottabretti
aka barnakerru
aka (< bílunum >) um vegleysur
aka dráttarvél
aka/geta ekið (< nokkurn >) bug á < keisarann, Frakka >
aka greitt
aka/hafa (< lengi >) ekið höllu
aka heilum vagni heim
aka hjólasleða
aka hjólbörum
aka < honum, henni> á bug
aka höllu fyrir < honum, henni >

10 orðasambönd með lýsingarorðinu feitur:

feit staða
feit steik
feitt embætti
feitt ket
feitt kjöt
feitt letur
feitt og holdasamt < fé >
hafa ekki um feitt að sleikja
vera feitur og líkamaþungur
vera feitur og pattaralegur

Orðasamböndin hafa nokkur séreinkenni sem sjá má á dæmunum að ofan. Meðal þeirra má nefna að svigaliðir koma mjög oft fyrir í þeim. Innihald sviganna er þá valfrjáls viðbót við orðasambandið sem þó er dæmigerð fyrir viðkomandi orðanotkun. Dæmi um slíkan svigalið er í sambandinu *aka (< bílunum >) um vegleysur*, þar sem hægt er að sleppa sviganum og hafa styttra form á sambandinu: *aka um vegleysur.*

Annað einkenni á orðasambandaskránni er að tilbrigði í tíðum sagna eru sýnd með skástriki á milli, eins og til dæmis í sambandinu *aka/hafa (< lengi >) ekið höllu*. Þetta er gert til að sögnin komi fram í

flettimynd sinni, nafnhætti, og til að notkun hennar birtist einnig á sem eðlilegastan hátt í orðasambandinu.

Í fyrstu tilraun til að marka orðasambönd voru notuð 2296 sambönd úr orðasambandaskránni sem öll byrjuðu á bókstafnum *a-*, án tillits til einstakra lykilorða innan sambandanna og orðflokks þeirra. Með **lykilorði** er átt við orð eins og *aka* og *feitur* í dæmunum að ofan, orð sem eru þungamiðja sambandanna eða í það minnsta mikilvæg orð innan þeirra. Það varð þó fljótlega ljóst að niðurstöður yrðu markvissari með því að miða mörkunina við ákveðin lykilorð fremur en stafrófsröðina sem orðasamböndin röðuðust eftir.

Í tilrauninni sem hér er til umræðu voru mörkuð 54 orðasambönd með sögninni *aka,* 29 orðasambönd með sögninni *afla,* 75 orðasambönd með lýsingarorðinu *feitur* og 46 orðasambönd með lýsingarorðinu *glaður.* Í 6. kafla verður litið á þessi orðasambönd og afdrif þeirra eftir mörkunarferlið.

3 Markari og mörkun

3.1 Markarinn

Þjálfun markara fyrir íslensku var tungutækniverkefni á vegum menntamálaráðuneytisins sem var unnið á Orðabók Háskólans frá haustinu 2002 fram í febrúar 2004 (sjá Eirík Rögnvaldsson o.fl. 2002). Til er greinargóð lýsing á verkefninu hjá Sigrúnu Helgadóttur (2004a og 2005) en hún sá um að þjálfa og prófa slík forrit fyrir íslensku. Fimm mismunandi markarar voru þjálfaðir fyrir íslenskan texta og reyndist svonefndur **TnT-markari** gefa besta raun. Markaranum var frá upphafi ætlað að vera fyrri hluti stærra verkefnis, og þegar þetta er ritað er verið að vinna að framhaldsverkefni hans sem ber nafnið **Mörkuð íslensk málheild** (sjá Sigrúnu Helgadóttur 2004b). Það felst í því að búa til markað textasafn (nefnt *tagged corpus* á ensku) sem hefur að geyma fullmarkaða íslenska nútímatexta, en slík textasöfn eru til víða erlendis.

Við mörkun orðasambandanna lá beinast við að nota það markaraforrit sem hafði gefið bestan árangur þegar því var beitt á venjulegan, samfelldan texta, það er að segja TnT-markarann. En þar sem orðasamböndin í þessari tilraun eru ekki samfelldur texti þótti ástæða til að prófa einnig aðra tegund markara því ekki var hægt að vita fyrir-

fram hver árangurinn af mörkun orðasambanda yrði og hvort hann lyti sömu lögmálum og þegar venjulegur texti var markaður.

Markarinn **fnTBL** var því einnig látinn spreyta sig á orðasamböndunum. Í stuttu máli sagt gaf fnTBL ekki sérlega góða raun. Mikill munur var á hæfni þessara tveggja markara til að greina orðasamböndin rétt, miklu meiri en þegar þeir voru látnir marka samfelldan texta.[2]

Markarinn fnTBL var aðeins notaður fyrst í stað til reynslu. Við greiningu 55 orðasambanda sem byrja á bókstafnum *a-* gerði fnTBL villur í 28 orðasamböndum þar sem TnT gerði aftur á móti villur í 6 orðasamböndum. Það samsvarar því að á móti hverju rangt greindu orðasambandi hjá TnT voru 4.67 rangt greind sambönd hjá fnTBL.

Ástæðan fyrir þessum mikla árangursmun er væntanlega sú að fnTBL-markarinn þarf talsvert meira textasamhengi en TnT til að geta markað rétt og í orðasamböndunum er samhengið oft mjög lítið (orðasamböndin eru oft mjög stutt, stundum eru þau aðeins tvö orð eins og til dæmis *glaða sólskin* og *afla matfanga*). Vegna þessa slaka árangurs fnTBL-markarans var horfið frá því að nota hann og ákveðið að nota eingöngu TnT-markarann við mörkun orðasambandanna.

3.2 Mörkun

Þegar texti er markaður fær hvert orð greiningarstreng sem fylgir því. Við þjálfun markaranna fyrir íslenskan texta var tekið mið af þeirri greiningu sem var notuð í *Íslenskri orðtíðnibók* sem kom út árið 1991 (Jörgen Pind o.fl. 1991). Markaskráin var einnig sú sama og í Orðtíðnibókinni, en með **markaskrá** er átt við lista yfir alla greiningarstrengi (mörk) sem koma fyrir í tilteknu mörkuðu textasafni (sbr. Sigrúnu Helgadóttur 2004a:60). Greiningarstrengurinn samanstendur af skammstöfunum þar sem fyrsti bókstafurinn táknar orðflokk og næstu stafir á eftir tákna málfræðiþætti sem eru breytilegir eftir orðflokki. Taka má dæmi til að lýsa þessu. Þegar orðasambandið *aka heilum vagni heim* er markað fæst eftirfarandi greining orðanna:

aka sng heilum lkeþsf vagni nkeþ heim aa

Sé lesið úr greiningarstrengjum orðanna stendur þetta:

[2]Hjá Sigrúnu Helgadóttur (2004a:61) kemur fram að TnT-markarinn náði 98.14% nákvæmni en fnTBL 97.25% nákvæmni þegar markararnir voru látnir greina orðin í textum Orðtíðnibókarinnar, og var þá miðað við einfaldasta mögulega greiningarstreng, þ.e. eingöngu orðflokkinn.

aka sng	sögn í nafnhætti, germynd
heilum lkeþsf	lýsingarorð í karlkyni, eintölu, þágufalli, sterkri beygingu, frumstigi
vagni nkeþ	nafnorð í karlkyni, eintölu, þágufalli
heim aa	atviksorð

Við mörkunina slitnar textinn í sundur og verður að lista þar sem eitt orð er í hverri línu. Mörkunarferlinu er lýst nánar í 4. kafla hér á eftir.

TnT-markarinn var þjálfaður á textasafni Orðtíðnibókarinnar. Hægt er að mæla nákvæmni mörkunar með fjölda réttra greiningarstrengja miðað við heildarfjölda þeirra. Sé aðeins tekið mið af orðflokki en ekki öðrum hlutum greiningarstrengsins, t.d. kyni, tölu og falli nafnorða, náði TnT-markarinn að meðaltali 98,14% nákvæmni við mörkun venjulegra texta (Sigrún Helgadóttir 2004a:61). Til samanburðar mörkuðust orðasambönd með sögninni *aka* með 95,07% nákvæmni (miðað við orðflokk eingöngu en þó með fallstjórnarmerkingu forsetninga og falli persónufornafna og afturbeygða fornafnsins), og orðasambönd með lýsingarorðinu *glaður* mörkuðust með 96,96% nákvæmni.

4 Ferli mörkunar

Mörkun texta fer þannig fram að gefin er skipun sem setur af stað mörkunarferlið og skilar það síðan niðurstöðuskrám. Sökum þess að orðasambönd eru ekki venjulegur texti þarf að gera sérstakar ráðstafanir til að mörkunin skili nothæfum niðurstöðum. Fyrst þarf að undirbúa textann, síðan er markaranum beitt á hann og loks er markaði textinn lagaður til svo að unnt sé að lesa úr honum niðurstöðurnar með hægu móti. Mörkun orðasambanda þarf því að fara fram í fimm til sex skrefum.

1. Orðasamböndin eru snyrt og einfölduð.
2. Orðasamböndin eru aðgreind hvert frá öðru með tákninu #.
3. Orðasamböndin eru mörkuð með TnT-markaranum sem skilar þeim sem lóðréttum lista.
4. Orðasamböndin eru færð aftur til fyrra horfs svo að hvert orðasamband sé í sérlínu.
5. Mörkin eru skilin frá orðunum og flutt aftast í línurnar.
6. Mörkin eru einfölduð eftir því sem ástæða þykir til.

1. Orðasamböndin er hægt að einfalda á tvennan hátt. Fyrsta skrefið er að taka út alla svigaliði, enda tákna þeir að innihald þeirra sé valfrjáls viðbót við orðasambandið.

Næsta skref er að huga að tilbrigðum í tíðum sagna (með skástriki á milli tilbrigða) eins og þau koma fram t.d. í orðasambandinu *aka/hafa (< lengi >) ekið höllu*. Þegar slík sambönd eru mörkuð kemur vel til greina að einfalda þau svo að aðeins fyrra tilbrigðið standi eftir. Ef sambandið *aka/hafa (< lengi >) ekið höllu* er einfaldað til fulls, er svigaliðurinn fyrst fjarlægður svo eftir stendur *aka/hafa ekið höllu*; því næst er sagnmyndin einfölduð í *aka höllu*. Við mörkun orðasambanda var bæði prófað að marka þau óbreytt sem og fullstytt, og hafði einföldun þeirra ekki sýnileg áhrif á nákvæmni mörkunarinnar. Í þessari tilraun var ákveðið að stytta orðasamböndin, en slíkt hlýtur að fara eftir aðstæðum og tilgangi hverju sinni.

2. Skilatáknið # er sett fremst og aftast í orðasambandið til að greina það frá öðrum samböndum. Það er nauðsynlegt vegna þess að hér er ekki um að ræða venjulegar setningar sem byrja á stórum staf og enda á greinarmerki — í orðasamböndunum koma greinarmerki önnur en komma mjög óvíða fyrir, og punktur kemur hvergi fyrir í þeim.

3. Nú er hægt að marka textann með TnT-markaranum og skilar hann niðurstöðum í sérstakar skrár. Ein þeirra hefur að geyma markaðan texta þar sem eitt orð er í línu ásamt markinu:

```
aka         sng
á           ao
hundasleða  nkeo
```

4. Til að fullvinna efnið þarf nokkur skref til viðbótar. Færa þarf orðasamböndin aftur til fyrra horfs svo að hvert samband sé í sérlínu. Á þessu stigi er hentugt að setja táknið _ á milli orðs og marks vegna síðari vinnslu:[3]

```
#  aka_sng á_ao hundasleða_nkeo #
#  aka_sng < honum_fpkeþ, henni_fpveþ > úr_aþ sporun-
   um_nkfþg #
```

Þarna eru þá orðasamböndin aftur komin í lárétta stöðu að viðbættum greiningarstrengjunum sem er skeytt aftan við hvert orð sambandsins.

5. Stundum getur verið hentugt að flytja alla greiningarstrengi orðasambandanna aftast í hverja línu, þ.e. á eftir orðasambandinu. Við það verður orðasambandið læsilegra og einnig auðveldar það áframhaldandi vinnslu á mörkuðu samböndunum, til dæmis ef ætlunin er að

[3] Þakkir fær Auður Rögnvaldsdóttir fyrir að búa til forrit sem gerir þetta tvennt.

láta tölvuna raða lista af orðasamböndum eftir greiningarstrengjunum. Þetta er búið að gera í eftirfarandi dæmum:

aka á hundasleða # _sng _ao _nkeo
aka < honum , henni > úr sporunum # _sng < _fpkeþ , _fpveþ > _aþ _nkfþg

Ólæsilegasti hluti textans, greiningarstrengirnir, eru nú allir komnir aftast í línurnar í réttri röð miðað við orðin í sambandinu.

6. Loks eru greiningarstrengirnir einfaldaðir með sérstakri skriftu. Þá er upplagt að nota tækifærið og gera skammstafanirnar gegnsærri og læsilegri. Í töflu 1 má sjá skammstafanir markanna eftir að þeim hefur verið breytt.

Tafla 1
Skammstafanir sem eru notaðar í einfaldaðri gerð markanna

no	nafnorð
lo	lýsingarorð
so	sögn
ao	atviksorð
st	samtenging, nafnháttarmerki
to	töluorð
fn	fornafn (annað en afturbeygt fornafn og persónufornafn)
fn-n	persónufornafn í nefnifalli
fn-a	afturbeygt eða persónufornafn í þolfalli
fn-d	afturbeygt eða persónufornafn í þágufalli
fn-g	afturbeygt eða persónufornafn í eignarfalli
fs-a	forsetning með þolfalli
fs-d	forsetning með þágufalli
fs-g	forsetning með eignarfalli

Orðasamböndin líta svona út eftir að búið er að einfalda greiningarstrengina. Þau eru orðin töluvert læsilegri:

aka á hundasleða # so fs-a no
aka < honum , henni > úr sporunum # so < fn-d , fn-d > fs-d no

5 Nákvæmni mörkunar og helstu takmarkanir

Vert er að hafa í huga að mörkun orðasambanda er krefjandi verkefni fyrir markann og er það fyrst og fremst vegna þess hve textasamhengið er stutt. TnT-markarinn vinnur á þann hátt að hann skoðar orðið sem hann er að greina ásamt tveimur orðum á undan því. Oft eru orðasamböndin aðeins tvö eða þrjú orð og er árangur hans við

mörkun þeirra þess vegna heldur verri en þegar hann er látinn marka samfelldari texta, eins og búast má við.

Það eru þrír þættir sem hafa mest áhrif á það hvernig tekst til með mörkun orðasambandanna: 1) markaraforritið sem er notað, 2) sú nákvæmni í greiningunni sem krafist er og 3) mörkun með eða án viðbótarorðasafns.

Markaraforritið

Eins og fram kom í kafla 3.1 var mikill munur á nákvæmni í greiningu milli þeirra tveggja markara sem voru prófaðir. Við mörkun orðasambanda gaf TnT-markarinn miklu betri niðurstöður en fnTBL-markarinn.

Nákvæmni greiningarstrengja

Því flóknari sem greiningarstrengir eru þeim mun meiri líkur eru á villum í þeim. Í *Íslenskri orðtíðnibók* er notuð stór markaskrá með 639 mismunandi greiningarstrengjum (sbr. Sigrúnu Helgadóttur 2005:258), en það er sama safn greiningarstrengja og það sem TnT-markarinn notar. Það er þó mjög bundið aðstæðum hversu nákvæma greiningu nauðsynlegt er að hafa og oft er alveg nóg að fá orðflokkinn greindan rétt. Heill greiningarstrengur lýsingarorðs er 6 stafir, greiningarstrengur sagnar í persónuhætti er sömuleiðis 6 stafir og nafnorðs 4–5 stafir. Nóg er að einn stafur í strengnum sé rangur til að markið teljist rangt.

Við mörkun orðasambandanna var miðað við að orðflokkurinn væri almennt nægilegur. Vegna séreinkenna orðasambandaskrárinnar var þó tekið með fall persónufornafna og afturbeygða fornafnsins, og fallstjórn forsetninga var einnig höfð með.

Viðbótarorðasafn

Það skiptir miklu máli hvort markað er með eða án viðbótarorðasafns (lexíkons). Þegar TnT-markarinn var þjálfaður var notast við orðasafn *Íslenskrar orðtíðnibókar*, en í því eru 31.876 orð í flettimynd sem koma fyrir í 59.358 orðmyndum (sjá Jörgen Pind o.fl. 1991:2 og Sigrúnu Helgadóttur 2004a:59). Það telst ekki vera stórt orðasafn. Orðasamböndin í skrá Orðabókar Háskólans eru eins og fyrr segir unnin upp úr ritmálssafni þeirrar stofnunar, og þótt þau séu mörg hver afar stutt er þar að finna auðugan orðaforða, eða 57.545 mismunandi lykilorð í flettimynd sinni samkvæmt talningu greinarhöfundar. Með því að nota aðeins orðasafn *Íslenskrar orðtíðnibókar* við mörkun orðasambandanna urðu niðurstöðurnar ekki sérlega góðar þar sem

mikið greindist af óþekktum orðum, sem markarinn gaf til kynna með sérstöku tákni. Lýsingarorð sem markarinn þekkti ekki greindi hann oft sem nafnorð og öfugt, auk þess sem hann gerði fleiri mistök í greiningu orðflokks.

Svo heppilega vill til að stórt viðbótarorðasafn er aðgengilegt. Það eru orðin sem beygð voru fyrir verkefnið *Beygingarlýsing íslensks nútímamáls*, sem var eitt af tungutækniverkefnum menntamálaráðuneytisins, unnið á Orðabók Háskólans. Í útgáfu 2.0 frá 2004 eru beygð 176.030 orð sem alls hafa 4.854.094 mismunandi beygingarmyndir, eins og kemur fram hjá Kristínu Bjarnadóttur, verkefnisstjóra Beygingarlýsingarinnar (2004:23–24).

Þegar markarinn var látinn nota viðbótarorðasafnið auk grunnorðasafnsins batnaði árangur hans stórlega. Allar niðurstöður úr mörkun sem birtar eru á þessum síðum voru fengnar með því að nota viðbótarorðasafnið.

Algengar villur

Algeng villa í greiningu orðasambandanna er sú að nafnháttur sagnar fremst í sambandinu greinist sem nafnorð eða lýsingarorð eins og sjá má á eftirfarandi dæmi. Hér eru greiningarstrengirnir hafðir eins og markarinn skilar þeim:

afla_nkeþ soðfisks_nkee

Þetta er röng greining, sögnin er greind sem nafnorð í karlkyni, eintölu, þágufalli.

afla_sng sér_fpkeþ fjár_nhee

Hér er greiningin hins vegar rétt, sögnin er greind sem sögn í nafnhætti.

Annar galli á markaranum er sá að hann á erfitt með að tengja forsetningu við fallorðið sem hún stýrir. Þegar greiningarstrengurinn er einfaldaður á þann hátt að aðeins orðflokkurinn er eftir kemur það ekki svo mjög að sök, en þegar það þarf nákvæmari greiningu er þetta óheppilegt. Dæmi um það þegar markarinn skilur ekki tengslin milli forsetningar og fallorðs má sjá í eftirfarandi orðasambandi:

þar_aa fór_sfg3eþ feitur_lkensf biti_nken í_aþ <Þjóðverja_nkfe-s>
(þar fór feitur biti í <Þjóðverja>)

Í orðasambandinu er *í* greint sem forsetning sem stýrir þágufalli (greining **aþ**). *Þjóðverja* er greint sem nafnorð í karlkyni, fleirtölu, eign-

arfalli, sérnafn (greining **nkfe-s**). Þótt ekki sé neinn vafi á því að forsetningin *í* stýri fallinu á nafnorðinu *Þjóðverja* virðist markarinn ekki geta tengt þetta tvennt saman. Ástæðan er að hluta til sú að orðið *Þjóðverja* er haft innan 'breytiliða' sem trufla samhengið frá sjónarmiði markarans, en þó kemur stundum fyrir að mörkuð fallstjórn forsetningar samsvarar ekki falli orðsins sem kemur beint á eftir, þótt það sé breytiliðalaust.

Ruglingur með fallstjórn forsetninga er algeng villa sem öllum mörkurum hættir til að gera, samkvæmt Sigrúnu Helgadóttur (2004a:59). Kafli 4 endaði á því að sýnd voru tvö orðasambönd þar sem búið var að einfalda greininguna. Þau voru þessi:

```
#   aka á hundasleða # so fs-a no
#   aka < honum , henni > úr sporunum # so < fn-d , fn-d > fs-d no
```

Fyrra dæmið að ofan, *aka á hundasleða*, fékk greininguna **so fs-a no**. Það er: sögn, forsetning með þolfalli, nafnorð. Þetta er ekki rétt greining, fallstjórn forsetningarinnar hefði átt að greinast þágufall en ekki þolfall þar sem orðið *hundasleða* er í þágufalli. Ljóst er að markarinn getur ekki vitað um hvort fallið er að ræða þar sem *hundasleða* er eins í þolfalli og þágufalli, og þar að auki getur forsetningin *á* stýrt báðum þessum föllum. Þarna hefði lengra samhengi ekki getað komið markaranum til hjálpar þar sem hann tekur ekki mið af mismunandi merkingu í samböndum eins og *aka á hundasleða* (þágufall) og *aka á ljósastaur* (þolfall), og orðin *hundasleða* og *ljósastaur* eru eins í þolfalli og þágufalli.

Þessi ruglingur á þolfalli og þágufalli er dæmigerð villa sem kemur stundum fyrir í mörkun orðasambandanna.

6 Mörkuð orðasambönd

Í þessum kafla eru fjórir listar með orðasamböndum, fyrst með sögnunum *aka* og *afla* og svo með lýsingarorðunum *feitur* og *glaður*. Búið er að einfalda samböndin á þann hátt sem lýst er í kafla 4: svigaliðir hafa verið teknir út og tilbrigði í sagnmyndum hafa verið einfölduð. Þar sem villa er í greiningunni er hún höfð með **feitu letri** og jafnframt er merkið ** sett á eftir sambandinu.

Nákvæmni mörkunar í orðasamböndunum er mæld á tvennan hátt. Í fyrsta lagi er miðað við orðasambandið sem heild og ef ein eða

fleiri villur koma fyrir í heildarmörkum orðanna er það talið rangt greint. Í öðru lagi er nákvæmnin metin eftir fjölda einstakra orða sem eru rétt greind, en það er sú aðferð sem Sigrún Helgadóttir (2004a og 2005) notar til að mæla nákvæmni mörkunar.

54 orðasambönd með sögninni aka:

1 aka léttan # so lo
2 aka **á** hundasleða # so **fs-a** no **
3 aka sér úr hlassastöðunum # so fn-d fs-d no
4 aka of þungu hlassi # so ao lo no
5 aka < vörunum > á hjólsleðum # so < no > fs-d no
6 aka hjólbörum # so no
7 aka hjólasleða # so no
8 aka < vatninu > **á** hestkerru # so < no > **fs-a** no **
9 aka < steypuefninu > á hestakerru # so < no > fs-d no
10 aka höllu fyrir < honum, henni > # so lo fs-d < fn-d, fn-d >
11 aka höllu # so lo
12 aka í hágír # so fs-d no
13 aka **greitt** # so **so** **
14 aka í fereykisvagni # so fs-d no
15 aka **á** þvottabretti # so **fs-a** no **
16 aka á völl # so fs-a no
17 að aka á völl # st so fs-a no
18 aka um vegleysur # so fs-a no
19 eiga heilum vagni heim að aka # so lo no ao st so
20 aka heilum vagni heim # so lo no ao
21 aka á tún # so fs-a no
22 < honum, henni > ekst < illa > í tauma # < fn-d, fn-d > so < ao > fs-a no
23 < allt > ekst í tauma # < fn > so fs-a no
24 < þetta > ekst < honum, henni > **í** tauma # < fn > so < fn-d, fn-d > **fs-d** no **
25 < sá uggur > ekst < honum, henni > **í** tauma # < fn no > so < fn-d, fn-d > **fs-d** no **
26 aka spölkorn # so no
27 aka < honum, henni > úr sporunum # so < fn-d, fn-d > fs-d no
28 akast úr sporunum # so fs-d no
29 aka sleðanum # so no
30 aka í sleða # so fs-d no
31 aka skrykkjótt # so lo
32 aka í skrautvagni # so fs-d no
33 aka seglum # so no
34 aka seglum eftir veðri # so no fs-d no

35 aka seglunum eftir veðri # so no fs-d no
36 aka seglunum eftir veðrinu # so no fs-d no
37 aka seglunum eftir vindinum # so no fs-d no
38 aka seglum eftir vindi # so no fs-d no
39 < honum, henni > er nærri ekið # < fn-d, fn-d > so ao so
40 aka mykju # so no
41 aka í móinn # so fs-a no
42 aka í léttivagni # so fs-d no
43 aka léttivagni # so no
44 < þetta > ekur < öllu > á kaldan klaka # < fn > so < fn > fs-a lo no
45 < **þeir,** þær, þau > akast á erindum um < þetta > < þangað til-S > # < **fn,** fn-n, fn-n > so fs-d no fs-a < fn > < ao tp > **
46 aka í drosju # so fs-d lo
47 aka dráttarvél # so no
48 aka í dagvagni # so fs-d no
49 láta ekki aka sér á bug # so ao so fn-d fs-a no
50 aka bug á < keisarann, Frakka > # so no fs-a < no, no >
51 aka < honum, henni > á bug # so < fn-d, fn-d > fs-a no
52 aka angri á bug # so no fs-a no
53 aka í bifreið # so fs-d no
54 aka barnakerru # so no

Orðasamböndin með *aka* markast þannig að af 54 samböndum greinast 47 rétt en villa kemur fyrir í 7 þeirra. Það samsvarar 87,04% nákvæmni sé miðað við heil orðasambönd en 95,07% nákvæmni sé miðað við einstök orð innan sambandanna.

29 orðasambönd með sögninni afla:

1 **afla** fiskjar # **no** no **
2 afla sér fæðslu # so fn-d no
3 afla fóðurforða # so no
4 afla sér fjár # so fn-d no
5 afla sér þekkingar # so fn-d no
6 afla sér viðar # so fn-d so
7 **afla** viðskiptasambanda # **no** no **
8 **afla** vetrarforða # **no** no **
9 afla sér veltufjár # so fn-d no
10 < þetta > **aflar** vanheilsu # < fn > **no** no ***
11 afla sér vanblessunar # so fn-d no
12 afla sér tiltrúar # so fn-d no
13 **afla sölva # no so** ***
14 afla soðningar # so no
15 **afla** soðmatar # **no** no **
16 **afla** soðfisks # **no** no **
17 **afla** til soðs # **no** fs-g no **

18 það aflast til soðs # fn-n so fs-g no
19 **afla** næringar sinnar # **no** no fn **
20 **afla** nýmetis # **no** no **
21 afla sér menntunar # so fn-d no
22 **afla** matfanga # **no** no **
23 afla sér fjár og mannvirðingar # so fn-d no st no
24 afla sér lífsuppeldis # so fn-d no
25 afla sér kristindómsþekkingar # so fn-d no
26 **afla** til kola # **no** fs-g no **
27 afla sér kennaramenntunar # so fn-d no
28 afla sér doktorsnafnbótar # so fn-d no
29 afla sér **bjargar** # so fn-d **so** ***

Af 29 orðasamböndum með *afla* markast 13 vitlaust en 16 rétt. Þetta er ekki góður árangur, aðeins 55,17% nákvæmni miðað við orðasambandið sem heild en 82,5% miðað við einstök orð. Það stafar einkum af því að markarinn ruglar saman sögninni *afla* og nafnorðsmyndinni *afla*, en þrjár mismunandi beygingarmyndir af karlkynsorðinu *afli* eru *afla* (ef orðið er ekki notað í fleirtölu). Hér þarf að grípa til eftirvinnslu sem felst í því að breyta marki orðs sem endar á *-a* í upphafi orðasambands úr nafnorði (no) í sögn (so).

Eftir þessa leiðréttingaraðgerð fremst í línu verða aðeins 3 villur eftir í mörkuninni og nákvæmnin mælist þá 89,65% miðað við orðasamböndin sem heild en 96,25% miðað við einstök orð.

Villurnar þrjár sem eftir eru eru þá þessar (merktar með 3 stjörnum): 1) í sambandinu < *þetta* > *aflar vanheilsu* er miðorðið greint sem nafnorð í stað sagnar. 2) Sambandið *afla sölva* er rangt greint því *sölva* er sagt vera sögn en ekki nafnorð, og 3) í sambandinu *afla sér bjargar* er síðasta orðið greint sem sögn í stað nafnorðs.

75 orðasambönd með lýsingarorðinu feitur:

1 < éta yfir sig þá sjaldan > hnífur < þeirra > kemur í feitt ket # < so fs-a fn-a ao ao > no < fn-e > so fs-a lo no
2 **feiti** kláði # **so** no **
3 feit jörð # lo no
4 feit leirjörð # lo no
5 feit **mjólk** # lo **lo** **
6 feit mold # lo no
7 feit moldarjörð # lo no
8 feit moldjörð # lo no
9 feit olía # lo no
10 feit sauðarföll # lo no

11 feit staða # lo no
12 feit steik # lo no
13 feitt embætti # lo no
14 feitt ket # lo no
15 feitt kjöt # lo no
16 feitt letur # lo no
17 feitt og holdasamt < fé > # lo st lo < no >
18 **feitt** og holdsamt < fé > # **so** st lo < no > **
19 feitt prestakall # lo no
20 feitt prófastsdæmi # lo no
21 feitt slátur # lo no
22 feitu níurnar # lo no
23 feitur biti # lo no
24 feitur dráttur # lo no
25 feitur eldishestur # lo no
26 feitur jarðvegur # lo no
27 feitur moldarjarðvegur # lo no
28 feitur naflastrengur # lo no
29 feitur og holdlaginn < uxi > # lo st lo < no >
30 finna ekki feitan gölt að flá < í íslenskum heimildum > # so ao lo no st lo < fs-d lo no >
31 finna hvar er feitt á stykkinu # so ao so lo fs-d no
32 finna < lítið > feitt á stykki # so < lo > lo fs-d no
33 < finna, sjá > það sem feitt er á stykkinu # < so, so > fn st lo so fs-d no
34 finna **það** sem **feitt** er á stykkinu # so **fn-n** st **so** so fs-d no **
35 flá ekki feitan kött # so ao lo no
36 flá feitan grís # so lo no
37 flá feitan gölt # so lo no
38 flá feitan gölt af < síldarkaupunum > # so lo no fs-d < no >
39 < fuglinn > er feitur og bústinn # < no > so lo st lo
40 **góna á** < hann, hana > eins og hundur á feitt ket # **no fs-d** < fn-a, fn-a > ao st no fs-a lo no **
41 **góna á** < hann, hana > eins og hundur á feitt kjöt # **no fs-d** < fn-a, fn-a > ao st no fs-a lo no **
42 hafa ekki feitan gölt að flá # so ao lo no st so
43 hafa ekki um feitt að sleikja # so ao fs-a lo st so
44 hafa < varla lengi > um feitt að sleikja # so < ao ao > fs-a lo st so
45 < hesturinn > er feitur og strykinn # < no > so lo st lo
46 < hér > er ekki feitan gölt að flá # < ao > so ao lo no st so
47 < hér > er ekki um feitt að sleikja # < ao > so ao fs-a lo st so
48 < hér > er **feitt** á stykkjunum # < ao > so **so** fs-d no **
49 hnífur < hans, hennar > kemst í feitt # no < fn-e, fn-e > so fs-a lo

50 hnífur < hans, hennar > kemur í feitt # no < fn-e, fn-e > so fs-a lo
51 < honum, henni > fellur **feitt** flesk í ketil # < fn-d, fn-d > so **so** no fs-a no **
52 kjósa **heldur** magra sætt en feitan prósess # so **so** lo no st lo no **
53 ríða feitum hesti < þaðan, frá þeim viðskiptum > # so lo no < ao, fs-d fn no >
54 < síldin > er feit á fiskinn # < no > so lo fs-a no
55 sjá < þar > feitan slag á borði # so < ao > lo no fs-d no
56 < ufsinn > er feitur á fisk # < no > so lo fs-a no
57 vera feitur og árlegur # so lo st lo
58 vera feitur og hjassalegur # so lo st lo
59 vera feitur og ístrumikill # so lo st lo
60 vera feitur og líkamaþungur # so lo st lo
61 vera feitur og pattaralegur # so lo st lo
62 vera feitur og sællegur # so lo st lo
63 vera feitur og þriflegur # so lo st lo
64 vera feitur svoli # so lo no
65 það er ekki feitan gölt að flá # fn-n so ao lo no st so
66 það er ekki feitt að sleikja # fn-n so ao lo st so
67 það er ekki um feitt að sleikja # fn-n so ao fs-a lo st so
68 það er < sjaldnast > feitum hesti heim að ríða # fn-n so < ao > lo no ao st so
69 < þar > er af feitum bita að klípa # < ao > so fs-d lo no st so
70 < þar > er ekki feitan gölt að flá # < ao > so ao lo no st so
71 < þar > er feitt í búi # < ao > so lo fs-d no
72 þar fór feitur biti **í** < Þjóðverja > # ao so lo no **fs-d** < no-s > **
73 < þar, þarna > er **feitt** á stykkinu # < ao, ao > so **so** fs-d no **
74 < þessir atburðir > eru feitir til frásagna # < fn no > so lo fs-g no
75 þrýstinn og feitur < sauður > # lo st lo < no >

Orðasamböndin með *feitur* markast þannig að af 75 orðasamböndum voru 11 rangt greind. Það samsvarar 85,33% nákvæmni í mörkun sé miðað við sambandið í heild. Alls fengu 14 orð ranga greiningu og sé miðað við einstök orð í textanum er nákvæmnin 95,88%.

46 orðasambönd með lýsingarorðinu glaður:

1 gera sér einn glaðan dag # so fn-d fn lo no
2 gera sér glaðan dag # so fn-d lo no
3 glaða sólskin # lo no
4 glaða tunglsljós # lo no
5 glaður eldur # lo no
6 glatt sólskin # lo no
7 glatt tunglskin # lo no
8 glatt veður # lo no
9 gleðjast með glöðum # so fs-d lo

10 glöð birta # lo no
11 glöð stjarna # lo no
12 góðan daginn, glaðan haginn # lo no, lo no
13 grípa glaðan **á** < öllu er getur glætt vonir þeirra um betri umskipti > # so lo **fs-a** < fn so so so no fn-e fs-a lo no > **
14 < hér > **brennur** < ekki lengur > glatt ljós í kolu # < ao > **no** < ao ao > lo no fs-d no **
15 með < **þeim** > brennur ekki glatt ljós í kolu < á síðustu árum > # fs-d < **fn** > so ao lo no fs-d no < fs-d lo no > **
16 < mér > er **glatt** í geði # < fn-d > so **so** fs-d no **
17 sjá < varla nokkurn > glaðan dag < upp frá þeim tíma > # so < ao fn > lo no < ao fs-d fn no >
18 < taka **þeim** kostum > með glöðu geði # < so **fn-d** no > fs-d lo no **
19 vera brosmildur og glaður # so lo st lo
20 vera glaður af víni # so lo fs-d no
21 vera glaður í anda # so lo fs-d no
22 vera glaður og fagnaðarfullur # so lo st lo
23 vera glaður og glaummikill # so lo st lo
24 vera glaður og góðfelldur # so lo st lo
25 vera glaður og góðkátur # so lo st lo
26 vera glaður og góðsinnaður # so lo st lo
27 vera glaður og hreifur # so lo st lo
28 vera glaður og kátur # so lo st lo
29 vera glaður og léttlyndur # so lo st lo
30 vera glaður og líflegur # so lo st lo
31 vera glaður og rósamur # so lo st lo
32 vera glaður og skemmtinn # so lo st lo
33 vera glaður og skrafhreifinn # so lo st lo
34 vera glaður og spaugsamur # so lo st lo
35 vera glaður og svipléttur # so lo st lo
36 vera glaður og viðmótsgóður # so lo st lo
37 vera glaður og viðmótsþýður # so lo st lo
38 vera glaður yfir < þessu > # so lo fs-d < fn >
39 vera himinljómandi glaður # so ao lo
40 vera hjartanlega glaður # so ao lo
41 vera hýrlyndur og glaður # so lo st lo
42 vera innilega glaður # so ao lo
43 vera óumræðilega glaður # so ao lo
44 verða glaður við að < þessum bágindum létti af > # so lo fs-a st < fn no so fs-d >
45 það er glaða sólskin # fn-n so lo no
46 < þar > **brennur** glatt ljós í kolu # < ao > **no** lo no fs-d no **

Orðasamböndin með *glaður* markast þannig að af 46 orðasamböndum voru 6 rangt greind. Það samsvarar 86,94% nákvæmni í mörkun sé miðað við sambandið í heild, en sé miðað við einstök orð er nákvæmnin 96,96%.

7 Röðun sambanda og kortlagning orða

Eftir að orðasamböndin hafa verið mörkuð má fara að huga að því að nota mörkunina á einhvern hátt. Það má til dæmis prófa að raða orðasamböndunum, ekki í hefðbundna stafrófsröð heldur eftir greiningarstrengjunum, þ.e. raða þeim eftir því sem kemur á eftir tákninu #. Þetta er gert til þess að reyna að fá fram eins konar kortlagningu af helstu setningargerðum með ákveðnu orði.

7.1 Orðasambönd með ‘feitur’

Fyrir neðan eru 60 orðasambönd með lýsingarorðinu *feitur.* Búið er að leiðrétta villur í mörkuninni og raða samböndunum eftir greiningarstrengjunum.

<þar, þarna> er feitt á stykkinu	# <ao, ao> so lo fs-d no
<hér> er ekki um feitt að sleikja	# <ao> so ao fs-a lo st so
<hér> er ekki feitan gölt að flá	# <ao> so ao lo no st so
<þar> er ekki feitan gölt að flá	# <ao> so ao lo no st so
<þar> er af feitum bita að klípa	# <ao> so fs-d lo no st so
<hér> er feitt á stykkjunum	# <ao> so lo fs-d no
<þar> er feitt í búi	# <ao> so lo fs-d no
þar fór feitur biti í <Þjóðverja>	# ao so lo no fs-a <no-s>
<honum, henni> fellur feitt flesk í ketil	# <fn-d, fn-d> so lo no fs-a no
<þessir atburðir> eru feitir til frásagna	# <fn no> so lo fs-g no
það er <sjaldnast> feitum hesti heim að ríða	# fn-n so <ao> lo no ao st so
það er ekki um feitt að sleikja	# fn-n so ao fs-a lo st so
það er ekki feitan gölt að flá	# fn-n so ao lo no st so
það er ekki feitt að sleikja	# fn-n so ao lo st so
feit moldarjörð	# lo no
feit moldjörð	# lo no
feit mold	# lo no
feit olía	# lo no

feit staða	# lo no
feit steik	# lo no
feitt embætti	# lo no
feitt ket	# lo no
feitt kjöt	# lo no
feitt letur	# lo no
feitt prestakall	# lo no
feitt prófastsdæmi	# lo no
feitur eldishestur	# lo no
feitur jarðvegur	# lo no
feitt og holdasamt <fé>	# lo st lo <no>
feitt og holdsamt <fé>	# lo st lo <no>
feitur og holdlaginn <uxi>	# lo st lo <no>
þrýstinn og feitur <sauður>	# lo st lo <no>
hnífur <hans, hennar> kemst í feitt	# no <fn-e, fn-e> so fs-a lo
hnífur <hans, hennar> kemur í feitt	# no <fn-e, fn-e> so fs-a lo
<síldin> er feit á fiskinn	# <no> so lo fs-a no
<ufsinn> er feitur á fisk	# <no> so lo fs-a no
<fuglinn> er feitur og bústinn	# <no> so lo st lo
hafa <varla lengi> um feitt að sleikja	# so <ao ao> fs-a lo st so
hafa ekki um feitt að sleikja	# so ao fs-a lo st so
flá ekki feitan kött	# so ao lo no
sjá <þar> feitan slag á borði	# so <ao> lo no fs-d no
finna ekki feitan gölt að flá <í íslenskum heimildum>	# so ao lo no st lo <fs-d lo no>
kjósa heldur magra sætt en feitan prósess	# so ao lo no st lo no
hafa ekki feitan gölt að flá	# so ao lo no st so
finna hvar er feitt á stykkinu	# so ao so lo fs-d no
finna það sem feitt er á stykkinu	# so fn-a st lo so fs-d no
sjá það sem feitt er á stykkinu	# so fn-a st lo so fs-d no
góna á <hann, hana> eins og hundur á feitt ket	# so fs-a <fn-a, fn-a> ao st no fs-a lo no
góna á <hann, hana> eins og hundur á feitt kjöt	# so fs-a <fn-a, fn-a> ao st no fs-a lo no
finna <lítið> feitt á stykki	# so <lo> lo fs-d no
flá feitan grís	# so lo no
flá feitan gölt	# so lo no
vera feitur svoli	# so lo no
ríða feitum hesti <þaðan, frá þeim við-skiptum>	# so lo no <ao, fs-d fn no>

flá feitan gölt af <síldarkaupunum>	# so lo no fs-d <no>
vera feitur og árlegur	# so lo st lo
vera feitur og hjassalegur	# so lo st lo
vera feitur og pattaralegur	# so lo st lo
vera feitur og sællegur	# so lo st lo
vera feitur og þriflegur	# so lo st lo

Næsta skref er að sleppa orðasambandinu svo greiningin ein standi eftir. Línur sem hafa eins greiningu eru aðeins sýndar einu sinni, og talan á undan markastrengnum táknar hér tíðni. Við þetta hefur línunum fækkað úr 60 í 32:

```
  1  < ao, ao > so lo fs-d no
  1  < ao > so ao fs-a lo st so
  2  < ao > so ao lo no st so
  1  < ao > so fs-d lo no st so
  2  < ao > so lo fs-d no
  1  ao so lo no fs-a < no-s >
  1  < fn-d, fn-d > so lo no fs-a no
  1  < fn no > so lo fs-g no
  1  fn-n so < aoe > lo no ao st so
  1  fn-n so ao fs-a lo st so
  1  fn-n so ao lo no st so
  1  fn-n so ao lo st so
 14  lo no
  4  lo st lo < no >
  2  no < fn-e, fn-e > so fs-a lo
  2  < no > so lo fs-a no
  1  < no > so lo st lo
  1  so < ao ao > fs-a lo st so
  1  so ao fs-a lo st so
  1  so ao lo no
  1  so < ao > lo no fs-d no
  1  so ao lo no st lo < fs-d lo no >
  1  so ao lo no st lo no
  1  so ao lo no st so
  1  so ao so lo fs-d no
  2  so fn-a st lo so fs-d no
  2  so fs-a < fn-a, fn-a > ao st no fs-a lo no
  1  so < lo > lo fs-d no
  3  so lo no
  1  so lo no < ao, fs-d fn no >
  1  so lo no fs-d < no >
  5  so lo st lo
```

Í listanum að ofan má sjá á tíðnitölunni hvaða setningargerð með lýsingarorðunu *feitur* er algengust í orðasambandaskránni, þ.e. **lo no** (lýsingarorð nafnorð). Fjórtán orðasambönd af þeim sextíu sem hér var raðað hafa þessa setningargerð, en fulltrúi hennar er til dæmis sambandð *feitt prófastsdæmi*.

Næstalgengasta setningargerðin er **so lo st lo** (sögn lýsingarorð samtenging lýsingarorð) sem kemur fimm sinnum fyrir, og er það til dæmis orðasambandið *vera feitur og sællegur.*

Þriðja algengasta mynstrið er **lo st lo < no >** sem kemur fjórum sinnum fyrir. Um er að ræða setningargerðina **lýsingarorð samtenging lýsingarorð < nafnorð >** þar sem nafnorðið er innan breytiliða. Dæmi um slíkt orðasamband er *þrýstinn og feitur < sauður >.*

7.2 Orðasambönd með 'aka'

Í 6. kafla voru birt 54 orðasambönd með sögninni *aka.* Nú verður litið aftur á þessi sömu sambönd eftir að búið er að leiðrétta villurnar í mörkuninni og raða þeim á sama hátt og samböndunum með *feitur,* þ.e. eftir greiningarstrengnum. Til að einfalda samböndin enn meira er breytiliðurinn < *honum, henni* > styttur í < *honum* >.

1 < honum > ekst < illa > í tauma # < fn-d > so < ao > fs-a no
2 < honum > er nærri ekið # < fn-d > so ao so
3 < sá uggur > ekst < honum > í tauma # < fn no > so < fn-d > fs-a no
4 < þeir > akast á erindum um < þetta > < þangað til-S > # < fn-n > so fs-d no fs-a < fn > < ao _tp >
5 < allt > ekst í tauma # < fn > so fs-a no
6 < þetta > ekst < honum > í tauma # < fn > so < fn-d > fs-a no
7 < þetta > ekur < öllu > á kaldan klaka # < fn > so < fn > fs-a lo no
8 aka of þungu hlassi # so ao lo no
9 láta ekki aka sér á bug # so ao so fn-d fs-a no
10 aka < honum > á bug # so < fn-d > fs-a no
11 aka < honum > úr sporunum # so < fn-d > fs-d no
12 aka sér úr hlassastöðunum # so fn-d fs-d no
13 aka á tún # so fs-a no
14 aka á völl # so fs-a no
15 aka í móinn # so fs-a no
16 aka um vegleysur # so fs-a no
17 aka í drosju # so fs-d lo

18 aka á hundasleða # so fs-d no
19 aka á þvottabretti # so fs-d no
20 aka í bifreið # so fs-d no
21 aka í dagvagni # so fs-d no
22 aka í fereykisvagni # so fs-d no
23 aka í hágír # so fs-d no
24 aka í léttivagni # so fs-d no
25 aka í skrautvagni # so fs-d no
26 aka í sleða # so fs-d no
27 akast úr sporunum # so fs-d no
28 aka greitt # so lo
29 aka höllu # so lo
30 aka léttan # so lo
31 aka skrykkjótt # so lo
32 aka höllu fyrir < honum > # so lo fs-d < fn-d >
33 aka heilum vagni heim # so lo no ao
34 eiga heilum vagni heim að aka # so lo no ao st so
35 aka barnakerru # so no
36 aka dráttarvél # so no
37 aka hjólasleða # so no
38 aka hjólbörum # so no
39 aka léttivagni # so no
40 aka mykju # so no
41 aka seglum # so no
42 aka sleðanum # so no
43 aka spölkorn # so no
44 aka angri á bug # so no fs-a no
45 aka bug á < keisarann , Frakka > # so no fs-a < no , no >
46 aka < steypuefninu > á hestakerru # so < no > fs-d no
47 aka < vatninu > á hestkerru # so < no > fs-d no
48 aka < vörunum > á hjólsleðum # so < no > fs-d no
49 aka seglum eftir veðri # so no fs-d no
50 aka seglum eftir vindi # so no fs-d no
51 aka seglunum eftir veðrinu # so no fs-d no
52 aka seglunum eftir veðri # so no fs-d no
53 aka seglunum eftir vindinum # so no fs-d no
54 að aka á völl # st so fs-a no

Þegar orðasamböndunum með sögninni *aka* að ofan er raðað eftir greiningarstrengnum sést vel kerfið sem þau mynda. Sem dæmi um það hvernig orðasamböndin hópa sig saman fá sambönd 49–53 öll sömu greininguna, en þessi fimm orðasambönd eru mismunandi tilbrigði af sama orðtakinu, *aka seglum eftir veðri*. Í línum 13–16 eru sambönd með forsetningarlið í þolfalli, og í línum 17–26 er sögnin með forsetningarlið í þágufalli. Þetta er einmitt aðalástæðan fyrir því að álitið var nauðsynlegt að taka með fallstjórn forsetninga í greiningar-

strengnum þótt hann væri einfaldaður niður í eintóman orðflokk alls staðar annars staðar — nema í sumum fornöfnum af þessari sömu ástæðu.

8 Mörkun orðabókartexta

Orðasambönd af því tagi sem hér eru til umræðu eru ekki venjulegur texti sem verður á vegi manna daglega. Þó minnir hann í mikilvægum atriðum á annan texta sem margir þekkja vel, þ.e. skýringartexta orðabóka. Þessar tvær textagerðir eiga ýmislegt sameiginlegt og má nefna eftirfarandi atriði:

- Textinn er ekki alltaf (og raunar sjaldnast) heilar setningar.
- Fremsta orðið stendur yfirleitt í flettimynd sinni, þ.e. sögn fremst stendur í nafnhætti (t.d. *aka á tún* í orðasambandaskránni) og nafnorð stendur í nefnifalli eintölu (t.d. *aðdráttarkraftur jarðar* í orðasambandaskránni).

Til að skýra þetta nánar má taka sögnina *aka* eins og hún er skýrð í *Íslenskri orðabók*, 3. útgáfu 2002.

Hér er aðeins birtur skýringartextinn við *aka* ásamt skýringartexta við orðasambönd og dæmi innan sagnarinnar. Það sem verið er að skýra er ekki sýnt, þ.e. sögnin sjálf og feitletruð og skáletruð sambönd og dæmi.

1 keyra, fara (flytja) á vagni, sleða e.þ.h.
2 nota hest sem dráttardýr og stjórna honum með aktaumum
3 flytja mykju (áður í kláfum á hesti) til áburðar á tún
4 þ.e. voru svo síðir á hestinum að þeir sópuðu snjónum með sér
5 hreyfa hægt, mjaka til
6 skaka sig örlítið, kiða sér
7 láta í minni pokann (fyrir e-m)
8 færa seglin (eftir vindstöðu og stefnu skipsins)
9 hegða sér eftir aðstæðum, leika tveim skjöldum
10 e-r er í klípu
11 hreyfast hægt, mjakast til
12 hrekja e-n
13 metast á, deila
14 malda í móinn
15 koma e-m í klípu
16 e-ð stöðvast hjá e-m

17 hreyfa sig smávegis
18 nudda e-m til verks
19 hörfa
20 sleppa við e-ð
21 e-r er alltaf jafn óheppinn
22 losna úr klípu

Tölurnar í fyrri dálknum eru settar hér til hagræðis en eru vitanlega ekki hluti af skýringartextanum. Þegar textinn var markaður voru sömu reglur notaðar og við mörkun orðasambanda, þ.e. orðflokkurinn var látinn nægja en þó tekið með fall persónufornafna og afturbeygða fornafnsins sem og fallstjórn forsetninga.

Hér að neðan má sjá hvernig mörkin líta út eftir að þau hafa verið einfölduð. Röng mörk eru með feitu letri. Línur þar sem röng mörk koma fyrir enda á **.

1 keyra, fara (flytja) á vagni, sleða e.þ.h. # so, so (so) fs-d no, no x
2 nota hest sem dráttardýr og stjórna honum með aktaumum # so no st no st so fn-d fs-d no
3 flytja mykju (áður í kláfum á hesti) til áburðar á tún # so no (ao fs-d no fs-d no) fs-g no fs-a no
4 **þ.e.** voru svo síðir á hestinum að þeir sópuðu snjónum með sér # **no** so ao lo fs-d no st fn-n so no fs-d fn-d **
5 **hreyfa hægt**, mjaka til # **no lo**, so fs-g **
6 skaka sig örlítið, kiða sér # so fn-a ao, so fn-d
7 láta í minni pokann (fyrir **e-m**) # so fs-a lo no (fs-d **no**) **
8 færa seglin (eftir vindstöðu og stefnu skipsins) # so no (fs-d no st no no)
9 hegða sér eftir aðstæðum, leika tveim skjöldum # so fn-d fs-d no, so to no
10 **e-r** er í klípu # **no** so fs-d no **
11 hreyfast hægt, mjakast til # so ao, so fs-g
12 hrekja **e-n** # so **lo** **
13 metast á, deila # so ao, so
14 malda **í** móinn # so **fs-d** no **
15 koma **e-m í** klípu # so **no fs-d** no **
16 **e-ð** stöðvast hjá **e-m** # **no** so fs-d **lo** **
17 hreyfa sig smávegis # so fn-a ao
18 nudda **e-m** til verks # so **no** fs-g no **
19 hörfa # so
20 sleppa við **e-ð** # so fs-a **no** **
21 **e-r** er alltaf jafn óheppinn # **so ao ao lo** **
22 losna úr klípu # so fs-d n**e>**

Það blasir við eftir þessa mörkun að sá háttur sem hefur tíðkast í orðabókum að nota skammstafanir á óákveðna fornafninu *einhver: e-r, e-n* o.s.frv., er ekki skiljanlegur markaranum. Í þessum 22 línum kemur skammstöfun á *einhver* fyrir í ýmsum föllum níu sinnum og er markið í öllum tilvikum rangt. *Einhver* er 7 sinnum greint sem nafnorð og tvisvar sem lýsingarorð.

Auðvitað væri hægt að bæta árangur markarans með því að lengja þessar skammstafanir fyrir mörkun þótt slíkt fæli í sér óæskilegt og tafsamt inngrip í textann. Önnur og betri leið væri að bæta skammstöfuninni *e-r* í öllum myndum inn í viðbótarorðasafnið.

Alls koma fyrir 14 villur í greiningu þessa orðabókartexta, þar af eru fimm villur af öðru tagi en skammstafanir á *einhver.* Það verður að teljast ágætur árangur því eins og í orðasamböndunum er í þessum stuttu textum lítið setningarlegt samhengi fyrir markarann að styðjast við.

9 Lokaorð

Þessi tilraun til að greina orðasambönd málfræðilega náði til um 200 orðasambanda sem höfðu eitt þessara fjögurra lykilorða: *afla, aka, feitur* og *glaður*. Tilgangurinn var upphaflega sá að prófa nýtt tól, TnT-markarann, og athuga hvernig gengi að marka orðasambönd með honum því ekki hafði áður verið reynt að greina þetta sérstaka textaform á vélrænan hátt í íslensku.

Árangur mörkunarinnar varð ekki fjarri þeim árangri sem hefur fengist þegar valinn samfelldur texti er markaður þótt hann væri að vísu heldur verri eins og gert hafði verið ráð fyrir. Nefna má tölur til samanburðar: TnT-markarinn náði að meðaltali 98,14% nákvæmni við mörkun venjulegra texta (miðað við orðflokk eingöngu) en orðasambönd með sögninni *aka* mörkuðust með 95,07% nákvæmni (miðað við orðflokk eingöngu en þó með fallstjórnarmerkingu forsetninga og falli persónufornafna og afturbeygða fornafnsins). Orðasambönd með lýsingarorðinu *glaður* mörkuðust með 96,96% nákvæmni.

Tveir þættir hafa mest áhrif á árangur mörkunarinnar: markarinn sem er notaður og mörkun með eða án viðbótarorðasafns. Tveir markarar voru prófaðir á orðasamböndunum og gaf TnT-markarinn mun

betri niðurstöður en fnTBL-markarinn. Notað var stórt viðbótarorðasafn þegar orðasamböndin voru mörkuð því þegar minna orðasafn var notað (orðasafn Orðtíðnibókarinnar) varð árangur mörkunarinnar með öllu óviðunandi.

Villur sem koma oftast fyrir við mörkun orðasambanda eru í fyrsta lagi röng fallstjórn forsetninga og í öðru lagi ruglingur með orðflokk sagnorðs sem er fremst í sambandinu. Dæmi um seinna atriðið er *afla til soðs* þar sem *afla* er greint sem nafnorð í stað sagnar.

Prófað var að nota niðurstöður úr greiningunni til að hópa saman líkum orðasamböndum. Það var gert með því að raða orðasamböndunum með *feitur* og *aka* eftir greiningarstrengnum. Þetta leiddi í ljós ákveðin mynstur í setningargerðum viðkomandi orða, t.d. sást að algengasta mynstrið með *feitur* í orðasambandaskránni er **lo no** (lýsingarorð nafnorð) sem er m.a. að finna í orðasambandinu *feitt prófastsdæmi*.

Tilraunin náði einnig að dálitlu leyti til skýringartexta *Íslenskrar orðabókar* og var árangur þeirrar mörkunar svipaður og í orðasamböndunum þótt ekki væru gerðar nákvæmar mælingar á niðurstöðunum.

Ljóst er að mörkun orðasambanda gefur áhugaverðar niðurstöður. Það má líta á hana sem leið til að fá yfirlit um helstu setningargerðir orða, og eru þá sagnir sérstaklega athyglisverðar þar sem setningargerðir þeirra eru oft mjög margbreytilegar. Slík aðkoma að sagnakerfinu væri ólík þeirri sem fæst við það að skoða flóknar sagnir eins og þær eru settar fram í orðabókum, og vísast þar einkum til meðferðar sagna í *Íslenskri orðabók* (2002) og *Orðastað* Jóns Hilmars Jónssonar (2001). En efnið í orðasamböndunum getur einnig nýst við greiningu á setningargerðum orða af öðrum orðflokkum, hægt er að kortleggja orð af öllum orðflokkum með þessari aðferð, sbr. lýsingarorðið *feitur* sem fjallað hefur verið um.

Heimildir

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

Íslensk orðabók. 2002 (3. útgáfa). Ritstj. Mörður Árnason. Reykjavík: Edda.

Jón Hilmar Jónsson. 2001. *Orðastaður* (2. útgáfa). Reykjavík: JPV.

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

Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. Í: *Samspil tungu og tækni. Afrakstur tungutækniverkefnis menntamálaráðuneytisins.* Bls. 23–25.

Sigrún Helgadóttir. 2004a. Markari fyrir íslenska texta. Í: *Samspil tungu og tækni. Afrakstur tungutækniverkefnis menntamálaráðuneytisins.* Bls. 55–64.

Sigrún Helgadóttir. 2004b. Mörkuð íslensk málheild. Í: *Samspil tungu og tækni. Afrakstur tungutækniverkefnis menntamálaráðuneytisins.* Bls. 65–71.

Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. Í: *Nordisk Sprogteknologi 2004 – Nordic Language Technology.* Bls. 257–265.

Heimildir á vefnum

Orðasambandaskrá Orðabókar Háskólans: www.lexis.hi.is/osamb/osamb.pl

Upplýsingar um Orðasambandaskrá Orðabókar Háskólans: www.lexis.hi.is/osamb/info.pl

Abstract

The topic of this article is the electronic tagging of phrases, fixed expressions and idioms as found in the collection of Orðabók Háskólans (Institute of Lexicography). The tool used for this, called *TnT-tagger,* grammatically analyses the words contained within the phrases. For this experiment, ca. 200 phrases were used, centred on two verbs (*afla* and *aka*) and two adjectives (*feitur* and *glaður*). The tagging process is described as well as the outcome of the tagging and the errors which occurred. The results are measured. The experiment also includes the tagging of several definitions from a dictionary (*Íslensk orðabók*). It is shown how the tags produced can be used for sorting the phrases so that they fall into groups of syntactic patterns. It is argued that this method is useful for finding patterns in the syntax of the keywords in question and for evaluating the frequency of various constructions.

Keywords:
phrases, idioms, grammatical analysis, PoS tagging, tagger

Þórdís Úlfarsdóttir
Orðabók Háskólans
Neshaga 16
107 Reykjavík
disa@lexis.hi.is