Eiríkur Rögnvaldsson

Textasöfn og setningagerð: greining og leit

1 Inngangur

Rannsóknir á flestum sviðum málvísinda byggjast með einum eða öðrum hætti á máldæmum. Setningafræðilegar rannsóknir eru þar engin undantekning. Það er hins vegar mjög misjafnt hvernig þessi dæmi eru fengin. Sumir setningafræðingar búa dæmi sín til sjálfir, en aðrir safna þeim úr rituðu eða töluðu máli. Til skamms tíma endurspegluðust mismunandi grundvallarviðhorf til viðfangsefnisins í þessum mismunandi uppsprettum dæma, en nú eru mörkin þarna á milli að verða óskýrari.

Þessi grein fjallar um setningafræðileg dæmi og leit að þeim í textasöfnum. Í byrjun er fjallað lítillega um texta og textasöfn sem setningafræðilegar heimildir. Bent er á að undanfarna áratugi hafa verið uppi mjög mismunandi viðhorf til gildis textasafna í setningafræðilegri umræðu og röksemdafærslu, en þann ágreining má að verulegu leyti rekja til mismunandi skoðana á því hvert viðfangsefni málfræðinnar sé; hvort hún eigi að lýsa málinu sjálfu (málbeitingunni) eða málkerfi málnotenda (málhæfninni). Einnig er fjallað nokkuð um margvíslegan vanda við túlkun þeirra upplýsinga sem textasöfn veita – ekki síst túlkun á þögn textanna um tilteknar setningagerðir.

Meginhluti greinarinnar fjallar um möguleika á setningafræðilegri dæmaleit í mismunandi greindum textasöfnum; hráum texta án nokk-

Orð og tunga 9 (2007), 51–73.

urra sérmerkinga, texta með beygingarlegri greiningu, og texta þar sem helstu setningarliðir og setningafræðileg hlutverk hafa verið greind. Sagt er frá tilraunum til að safna dæmum um tilteknar setningagerðir í íslenskum textasöfnum; einkum grunnskrám *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) og ÍSTAL-safninu (sjá Þórunni Blöndal 2005 og grein Ástu Svavarsdóttur (2007) í þessu hefti). Við slíka dæmasöfnun væri æskilegt að hafa aðgang að textum sem hafa verið greindir setningafræðilega, en engin söfn slíkra texta íslenskra eru til.

Nú er hins vegar hægt að greina íslenska texta beygingarlega (marka þá) á vélrænan hátt (sjá grein Sigrúnar Helgadóttur (2007) í þessu hefti). Í ljósi þess að verulegar setningafræðilegar upplýsingar felast í hinum málfræðilegu mörkum þótti ómaksins vert að athuga hvort hægt væri að nýta þau í setningafræðilegri leit. Í ljós kom að beygingarlega mörkunin gagnast mjög vel við leit að ýmsum setningagerðum, og gerir setningafræðilega dæmasöfnun margfalt fljótlegri og markvissari en áður. Í lokin er drepið stuttlega á ólokið verkefni þar sem unnið er að vélrænni hlutaþáttun (setningafræðilegri greiningu) sem gæti auðveldað setningafræðilega dæmaleit að mun.

2 Textasöfn sem heimild um setningagerð

Á fyrri hluta 20. aldar var talsvert gert af því í ýmsum löndum að nota stór textasöfn sem grundvöll mállýsinga. En viðhorfin til textasafna og megindlegra athugana á tungumáli gerbreyttust svo að segja á einni nóttu fyrir hálfri öld, eins og McEnery og Wilson benda á (1996:4):

> [...] we can pinpoint a discontinuity in the development of corpus linguistics fairly accurately in the late 1950s. After this period the corpus as a source of data underwent a period of almost total unpopularity and neglect. Indeed it is no exaggeration to suggest that as a methodology it was widely perceived as being intellectually discredited for a time.

Það sem olli þessum straumhvörfum var tilkoma málkunnáttufræði (generatífrar málfræði) Chomskys (1957). Chomsky gaf lítið fyrir gildi textaathugana en byggði málfræði sína þess í stað á máltilfinningu og

dómum málnotenda. Áhrif Chomskys voru mjög mikil og næstu áratugina þótti fæstum setningafræðingum ástæða til að leggjast í dæmasöfnun úr textum til að rökstyðja kenningar sínar, heldur bjuggu sjálfir til dæmi sín og dæmdu þau tæk eða ótæk. Þessi aðferð þykir vissulega enn góð og gild, en á seinni árum hafa menn aftur horfið til dæmasöfnunar úr textum og láta aðferðirnar vinna saman og bæta hvora aðra upp. Í þessum kafla er drepið á nokkrar forsendur þess ágreinings sem hefur verið um gildi textadæma og rætt sérstaklega um það hvernig mismunandi fræðilegar forsendur geta leitt til mismunandi túlkunar þess vitnisburðar sem textarnir gefa.

2.1 Heimild um málhæfni eða málbeitingu?

Ein meginröksemd Chomskys fyrir því að textasöfn væru gagnslaus í setningafræðilegri greiningu og röksemdafærslu var sú að þau væru ævinlega og óhjákvæmilega takmörkuð, endanleg, og tilviljanakennd (sjá t.d. Chomsky 1957:13-17). Auðvelt er t.d. að tilfæra ýmis dæmi um setningar og setningagerðir sem sjaldan eða aldrei finnast í textasöfnum, jafnvel mjög stórum, en málhöfum ber þó saman um að séu tækar. Þetta hefur oft verið notað sem rök fyrir því að málhæfnin sé að verulegu leyti meðfædd; menn geti ekki hafa lært slíkar setningar af öðrum, heldur hljóti að hafa einhverja meðfædda þekkingu á þeim reglum sem um þær gilda.

Skiptar skoðanir um þessi mál leiddu til hálfgerðs stríðs milli málkunnáttufræðinga (generatífista) og þeirra sem fengust við gagnamálfræði (corpus linguistics). Chomsky talaði víða óvirðulega um gagnamálfræði, og í ritum gagnamálfræðinga er að finna mörg og beitt skot á Chomsky og fylgismenn hans (sjá um þetta t.d. McEnery og Wilson 1996:4-17, 61-66 o.v.). Hér er þó rétt að halda því til haga að þarna er að verulegu leyti um sýndarágreining að ræða – meðvitað eða ómeðvitað. Menn voru nefnilega ekki að tala um sama hlutinn. Chomsky var að tala um málhæfni (competence) en gagnamálfræðingar skoða málbeitingu (performance) (sjá t.d. Chomsky 1965:4). Chomsky var sem sé að tala um málfræðina, málkerfið, en gagnamálfræðingar skoða afurð kerfisins – málið sjálft. Þarna á milli er flókin víxlverkun sem ekki hefur verið kortlögð til fulls, en meginatriðið er að báðar aðferðirnar eiga fullan rétt á sér og eru nauðsynlegar – en þær svara mismunandi spurningum.

Textasöfn eru þannig gagnleg til að finna ýmsar setningagerðir og átta sig á þeim. Það er t.d. hægt að nota þau, að vissu marki, til að úrskurða tiltekna setningagerð tæka. Það er hins vegar ekki hægt að nota þau til að úrskurða setningagerð ótæka. Þótt hún komi ekki fyrir í þeim textum sem við skoðum getur það verið tilviljun. Eðli málsins samkvæmt getur textasafn okkar aldrei innihaldið allar hugsanlegar setningar. Ef við erum að lýsa málinu (ekki málkerfinu) gerir þetta ekkert til. Textasafnið sem við höfum undir afmarkar þá viðfangsefni okkar, og ef tiltekin setningagerð kemur ekki fyrir í safninu er hún ekki hluti viðfangsefnisins og kemur okkur þess vegna ekkert við.

En ef við erum að lýsa málkerfinu sjálfu horfir málið öðruvísi við. Það málkerfi sem við lýsum á að gera okkur kleift að mynda allar málfræðilega tækar setningar en ekki aðrar. Þess vegna nægir okkur ekki að vita hvers konar setningar eru tækar – við þurfum líka að vita hvers konar setningar væru ótækar. Og því svarar textasafnið ekki – það er vitaskuld ekki hægt að takmarka mengið „tækar setningar“ við þær setningar sem fyrir koma í tilteknu safni, hversu stórt sem það er; „it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in his field work“, segir Chomsky (1957:15) og hnykkir enn á því síðar:

> [...] though “probability of a sentence (type)” is clear and well defined, it is an utterly useless notion, since almost all highly acceptable sentences (in the intuitive sense) will have probabilities empirically indistinguishable from zero and will belong to sentence types with probabilities empirically indistinguishable from zero. Thus the acceptable or grammatical sentences (or sentence types) are no more likely, in any objective sense of this word, than the others (Chomsky 1965:195).

2.2 Hversu marktækir eru textarnir?

Í samtímalegri setningafræði er hægt að snúa sig út úr þessum vanda með þeim einfalda hætti að spyrja málnotendur. Þá erum við ekki háð afmörkuðu textamengi, heldur getum búið til texta eftir þörfum, ef svo má segja, og borið þá undir málnotendur og fengið dóma þeirra

um hvort tiltekin setning sé tæk eða ekki. Þeir sem fást við sögulega setningafræði eiga aftur á móti ekki þessa útleið – þeir verða að reiða sig algerlega á textana (sjá umræðu um þetta hjá Eiríki Rögnvaldssyni 1998). Stundum hafa menn reynt að bæta sér það upp með einhverjum ráðum, eins og t.d. því sem kallað hefur verið „lögmál ónýttra tækifæra“ (Principle of missed opportunities) og er orðað svo:

(1) If a certain syntactic form is used regularly in a given function or type of context C in a living language L, and if F is absent in C at an earlier stage of the language, OL, then there is good reason to assume that F does not exist in OL (Faarlund 1990:17).

Þetta getur þó aðeins verið viðmið sem verður að beita af mikilli varfærni. Hvernig getum við t.d. fullyrt að eitthvað sé „absent [...] at an earlier stage of the language“ – hvernig skilgreinum við „language“ þarna? Við höfum ekki annað til að miða við en þá texta sem varðveittir eru eða við höfum aðgang að – og þeir eru ekki alltaf miklir. En við þurfum líka að gæta þess að skoða þá alla áður en við fullyrðum nokkuð, gæta þess að ekki komi eitthvað annað til sem geti valdið því að viðkomandi setningagerð finnst ekki á eldra málstigi, o.s.frv.

Hér á undan var sagt að hægt væri – að vissu marki – að nota textasöfn til að úrskurða tiltekna setningagerð tæka. En þar verður líka að hafa fyrirvara. Því fer nefnilega fjarri að allar setningar sem koma fyrir í textasöfnum séu tækar í raun og veru, þ.e. samræmist málkerfi flestra málnotenda. McEnery og Wilson (1996:13) hafa eftir Chomsky að allt að 95% allra segða (utterances) séu í raun málfræðilega ótækar. Þar er væntanlega miðað við talmál og hlutfallið örugglega mun lægra í rituðu máli – og McEnery og Wilson vilja líka meina að tala Chomskys sé alltof há. En jafnvel þótt við lítum framhjá mállýskumun er ljóst að í rituðu máli er nokkuð um setningar sem flestir myndu telja ótækar – setningar þar sem fyrir koma ýmiss konar pennaglöp, mistök í ritvinnslu, prent- og ásláttarvillur, einstaklingsbundið málfar, o.s.frv. Dæmi um þetta má sjá í eftirfarandi setningu af mbl.is:

(2) Alls voru um 179 tonn af hvalaúrgangi af þeim sem sjö langreyðum sem veiddust við landið í haust urðað að Fífilholtum á Mýrum. [Undirstrikun mín, ER]
http://www.mbl.is/mm/frettir/frett.html?nid=1245425

Þarna eru – að flestra mati, geri ég ráð fyrir – tvær villur; *sem* ofaukið og sambeygingu vantar. Þegar við lesum leiðréttum við þetta með

sjálfum okkur þegjandi og hljóðalaust, í samræmi við málkennd okkar. En hvenær getum við leyft okkur það? Hvað með þá sem fást við eldri málstig og geta ekki beitt eigin málkennd á textana? Verðum við að líta á allar setningar sem finnast í textum sem jafnréttháar? Yfirleitt gera menn það ekki í raun; „one must be ready to characterize certain unattested sentences as well-formed and some attested sentences as ill-formed", segir Lightfoot (1979:6.) En þarna eru menn vissulega á hálum ís, og oft getur verið freisting að láta fræðikenningar taka af sér ráðin; hafna setningum sem koma fyrir ef þær falla ekki að þeirri kenningu sem maður vinnur með, en gera ráð fyrir öðrum sem ekki finnast dæmi um, vegna þess að kenningin segir að þær ættu að geta komið fyrir.

2.3 Ályktanir af þögn textanna

Gott dæmi um það hvernig mismunandi fræðileg afstaða kemur fram í mismunandi túlkun á þögn textanna má taka úr lífseigri deilu um það hvort aukafallsfrumlög hafi verið til í fornu máli. Flestir málfræðingar fallast núorðið á að nútímaíslenska hafi aukafallsfrumlög, en margir halda því fram að tilkoma þeirra sé séríslensk þróun og samsvarandi liðir hafi ekki haft stöðu frumlags í fornu máli. Það hefur m.a. verið rökstutt með því að í fornmáli komi viðkomandi aukafallsliðir ekki fyrir í öllum sömu setningagerðum og í nútímamáli, t.d. ekki í setningum af þessu tagi:

(3) a. Ég vonast til að vanta ekki efni í ritgerðina. (Höskuldur Þráinsson 1979)
b. Hann vonast til að leiðast ekki. (Halldór Ármann Sigurðsson 1989)

Það skiptir að margra mati verulegu máli hvort einhver fornmálsdæmi finnist um slíkar setningar; Falk (1995:203) nefnir t.d. þessa setningagerð sem fullnaðarsönnun fyrir því að nútímaíslenska hafi aukafallsfrumlög. Mørck (1992) segist hafa leitað sérstaklega að dæmum á við (3) í fornum textum, en án árangurs, „så jeg meiner at vi får holde fast ved at lik-NP-stryking bare virker på nominativledd, inntil noe anna kan dokumenteres" (Mørck (1992:71). Ég hef hins vegar bent á (Eiríkur Rögnvaldsson 1996) að setningar á við (3) eru mjög sjaldgæfar í nútímamáli, að því er virðist – finnast ekki einu sinni þótt leitað

sé með *Google* í öllu því textamagni sem er að finna á netinu, og er auðvitað margfalt það sem til er á fornískensku.

Hver er þá niðurstaðan? Ég fæ ekki betur séð en hér geti hver trúað því sem hann vill. Það væri vissulega betra fyrir mig ef dæmi á við (3) fyndust í fornu máli, en ég get samt haldið því fram – með vísun til þess sem haft er eftir Chomsky hér að framan – að við því sé alls ekki að búast að þau finnist, og fjarvera þeirra segi ekkert um það hvort þau hafi verið tæk að fornu. Þeir sem vilja hafna tilvist aukafallsfrumlaga í fornu máli geta líka sagt: Fyrst engin dæmi af þessu tagi finnast þá höfum við enga sönnun fyrir því að þessi setningagerð hafi verið tæk í fornu máli, og meðan við höfum enga slíka sönnun getum við ekki leyft okkur að gera ráð fyrir aukafallsfrumlögum.

2.4 Breytt viðhorf til textadæma

Viðhorf margra málfræðinga til texta og textasafna hefur breyst á seinni árum. Nú þykir miklu eðlilegra en fyrir fáum árum að rökstyðja setningafræðilegar greiningar með dæmum úr töluðu eða rituðu máli. Hrint hefur verið af stað viðamiklum rannsóknarverkefnum til að kanna setningafræðilegan mállýskumun og safna setningafræðilegum dæmum, s.s. norræna verkefninu *Scandinavian Dialect Syntax* (*http://uit.no/scandiasyn*) og „dótturverkefnum“ þess, m.a. íslenska verkefninu *Tilbrigði í setningagerð* sem Höskuldur Þráinsson stýrir (sjá Ástu Svavarsdóttur 2006). Þetta hefði tæpast getað gerst fyrir 20 árum eða svo.

Að hluta til má skýra þessa þróun með því að máldæmi, einkum ritmálsdæmi, eru orðin mun auðfengnari en áður. Með tilkomu sístækkandi rafrænna textasafna er nú orðið auðvelt að safna fjölbreyttum textadæmum af ýmsu tagi. Vefurinn hefur svo gert mönnum kleift að komast í margs konar texta sem áður voru óaðgengilegir og einnig hafa þar orðið til nýjar textategundir sem margar hverjar standa nær talmálinu en hefðbundnu ritmáli. Leitarvélar á vefnum, eins og *Google* og *Embla*, hafa svo auðveldað mönnum dæmasöfnun úr þessum textum.

En þessi þróun býður líka hættunni heim og það verður að fara varlega við notkun og túlkun þeirra dæma sem aflað er með leit á netinu. Þótt þar finnist setningagerð sem menn þekktu ekki áður þarf það ekki að tákna að hún sé ný í málinu – eins gæti verið að hún hafi

lengi verið til, en tilheyri hins vegar málsniði sem ekki hefur áður verið notað í (aðgengilegu) ritmáli. Það þarf líka að hafa í huga að dæmi á netinu eru ekki endilega úr nútímamáli. Það er talsvert af fornum textum á netinu (t.d. allar *Íslendingasögur*, *Heimskringla* og *Fornaldarsögur Norðurlanda* hjá Netútgáfunni, *http://www.snerpa.is/net*). Þegar ég var að skoða samband fornafnanna *sjálfur* og *sinn* í fyrra fann ég á netinu allnokkur dæmi um *sjálfrar sinnar*; en þegar að var gáð reyndust þau flest vera úr eldri textum.

Eins þarf að gæta þess að talsvert er af málfræðigreinum á netinu og í þeim eru stundum dæmi sem annaðhvort eru beinlínis ótæk og eiga að vera það, eða koma sjaldan fyrir í venjulegum textum og eru því ekki marktæk sem dæmi um málnotkun. Hér að framan var því haldið fram að jafnvel í leit á netinu með *Google* fyndust engin dæmi á við (3) en það er ekki alveg rétt; í raun finnur *Google* fjögur dæmi um sambandið *vonast til að vanta ekki* og tvö dæmi um *vonast til að leiðast ekki*. En þegar dæmin eru skoðuð kemur í ljós að þau eru öll úr dæmasetningum málfræðinga.

2.5 Textasöfn í tungutækni

Hér er ekki ætlunin að gera ítarlega úttekt á kostum þess og göllum að nota textasöfn í setningafræðirannsóknum. Enginn vafi er á því að textasöfn geta komið að miklu gagni á því sviði, en hitt er jafnljóst að þau svara ekki öllum spurningum og nauðsynlegt er að gæta varúðar í túlkun þeirra. En þegar litið er á gildi textasafna frá sjónarhóli tungutækni er viðhorfið annað. Þar er sjónarhornið hagnýtt fremur en fræðilegt – ekki verið að leita upplýsinga um málkerfið, heldur greina textana og vinna úr þeim upplýsingar sem síðan er hægt að nota til að „hanna eða útbúa einhvern hugbúnað eða tæki sem nýtist mönnum í starfi eða leik“, eins og segir í skilgreiningu orðsins *tungutækni* í *Orðabanka Íslenskrar málstöðvar* (*http://herdubreid.rhi.hi.is:1026/wordbank/-search*). Þá skiptir ekki endilega máli hvers vegna tiltekin setningagerð kemur ekki fyrir – hvort það er vegna þess að hún er sjaldgæf, eða vegna þess að hún sé alls ekki hugsanleg í málinu; málfræðilega ótæk. Ef hún kemur ekki fyrir í stóru textasafni er ekki líklegt að mörg dæmi um hana komi fyrir í öðru safni sambærilegra texta. Því er ekki líklegt að hugbúnaður okkar eða tól þurfi að glíma við hana, nema þá í mjög litlum mæli. Jafnvel þótt setningagerðin komi fyrir, og hugbún-

aður okkar greini hana rangt eða alls ekki, hefur það þá ákaflega lítil áhrif á heildarframmistöðu búnaðarins.

3 Leit að setningagerðum í textasöfnum

Það er til lítils að koma upp safni af textum úr töluðu og rituðu máli ef ekki eru til aðferðir til að vinna úr þessum söfnum. Það þarf að vera hægt að leita í þeim að dæmum um tilteknar setningagerðir. Við þá leit má beita tveimur ólíkum aðferðum. Önnur er sú að lesa textana frá upphafi til enda og skrá dæmi úr þeim. Ókostur aðferðarinnar er vitanlega sá að hún er mjög seinleg, auk þess sem alltaf er hætta á að dæmi fari fram hjá lesandanum. Til skamms tíma var þetta þó eina aðferðin sem völ var á, en það hefur breyst á síðustu 20-25 árum með tilkomu rafrænna texta. Það væri því mikill kostur ef hægt væri að leita að dæmum á skipulegan hátt í tölvu. Bæði væri slík leit mjög fljótleg, og eins ætti hún að geta verið tæmandi – sé leitað á réttan hátt. Forsendur fyrir slíkri leit eru einkum tvær; að til séu tölvutækir textar, og að þeir séu málfræðilega greindir á þann hátt að hægt sé að leita að setningafræðilegum fyrirbærum.

Í þessum kafla er fjallað um mismunandi aðferðir við setningafræðilega dæmaleit í textum; frá einfaldri textaleit yfir í leit í beygingarlega mörkuðum textum, og að lokum um leit í setningafræðilega mörkuðum textum. Sagt er frá nokkrum tilraunum sem ég hef gert til að nýta beygingarlega mörkun í setningafræðilegum tilgangi og hafa gefið góða raun.

3.1 Textaleit

Einfaldasta form leitar er það sem öll ritvinnsluforrit bjóða upp á; að slá inn streng (eitt orð eða fleiri) og leita að honum, nákvæmlega eins og hann er ritaður. Smávægileg tilbrigði eru möguleg (t.d. að tilgreina hvort hástafir og lágstafir skipta máli), og stundum er hægt að nota algildisstafi (e. *wildcard characters*) til að leita að hvaða staf sem er. Í *Word* finnur b^?r til dæmis *bar*, *ber*, *byr*, *bor*, *bær*, *býr* o.s.frv. Í UNIX-stýrikerfinu er hægt að nota reglulegar segðir (e. regular expressions) við leitina og tilgreina þannig flókin leitarmynstur. Þannig finnur [iy]n[gk][^jiíeæ] strengina *ing*, *yng*, *ink* og *ynk*, en þó því aðeins að enginn stafanna *j*, *i*, *í*, *e*, *æ* komi næst á eftir. Ýmis sérhæfð texta-

vinnsluforrit bjóða upp á sérhæfðari möguleika. Í *WordCruncher* er t.d. hægt að leita að orðum sem koma fyrir nálægt hvort öðru, með tilgreindum hámarksstafafjölda á milli. Þar er líka hægt að hlaða orðum inn í bálka og leita að dæmum þar sem eitthvert orð úr öðrum bálkinum kemur fyrir í grennd við eitthvert orð úr hinum. Svona mætti lengi halda áfram að telja upp þá möguleika sem finnast í ýmsum forritum og auðvelda manni leitina.

Leit af þessu tagi er þó alltaf þeim takmörkunum háð að hún er bundin við orð. Það veldur því að erfitt er að nýta hana við leit að tilteknum setningagerðum – það er því aðeins hægt að unnt sé að tengja setningagerðir við ákveðin orð. Þannig er t.d. hægt að nýta slíka leit að vissu marki til að skoða afturbeygingu, með því að leita að myndum afturbeygða fornafnsins *sig/sér/sín*, og afturbeygða eignarfornafnsins *sinn/sín/sitt*. En jafnvel hér er þessi aðferð ófullnægjandi. Í fyrsta lagi vegna þess að sumar myndir afturbeygða fornafnsins og afturbeygða eignarfornafnsins falla saman við beygingarmyndir annarra orða (*sér* getur verið 3. pers. et. fh. nt. af *sjá*, og *sinn* getur verið hvorugkynsnafnorðið *sinn*). Því þarf að fara gegnum öll dæmin sem finnast við slíka leit og vinsa úr þeim. Það er seinlegt en ekki frágangssök. En til að fá góða mynd af notkun afturbeygingar þarf líka að skoða setningar þar sem afturbeygð fornöfn eru ekki notuð, heldur persónufornöfn. Þá vandast málið; því að persónufornöfn eru vitanlega notuð við miklu fjölbreyttari aðstæður. Þau eru svo algeng að það borgar sig ekki að leita að þeim; það væri svo mikið verk að vinsa úr leitarniðurstöðunum að það er alveg eins gott að lesa bara textann í heild.

Ég hef mikla reynslu af því að nota textaleit, einkum með hjálp *WordCruncher*, í setningafræðilegum rannsóknum á fornískenskum textum. Þeir textar sem ég vann með voru að mestu leyti ógreindir, þótt vissulega megi hafa nokkurt gagn af greiningunni í *Orðstöðulykli Íslendinga sagna* (Bergljót S. Kristjánsdóttir o.fl. 1996). Textaleitin hefur vissulega skilað ágætum árangri í mörgum tilvikum. Þannig hef ég t.d. leitað að dæmum um aukafallsfrumlög (Eiríkur Rögnvaldsson 1996) og boðháttarsagnir (Eiríkur Rögnvaldsson 2000). Í fyrra tilvikinu lá fyrir hverjar væru helstu sagnir sem kæmu til greina að tækju aukafallsfrumlög, og því var leitað að þeim; í því síðara var leitað að dæmum um boðháttarmyndir algengra sagna. Í báðum tilvikum dugði að finna (sem flest) dæmi; ekki var ætlunin að setja fram neina tölfræði á grundvelli þeirra, og því var aðferðin fullnægjandi.

En ég hef líka notað þessa aðferð til að skoða orðaröð í sagnlið og leita þar að tilteknum orðaraðarmynstrum (Eiríkur Rögnvaldsson 1994-95). Þar reyndi ég að sýna fram á að tiltekin mynstur kæmu fyrir innan flókinna sagnliða (einkum liða sem hafa að geyma tvær fallhátt-arsagnir og tvö andlög) en önnur ekki, og það væri kerfi í því hvað kæmi fyrir og hvað ekki. Í þessu tilviki var ekki einfalt að nota einstök orð í leitinni. Ég nýtti mér það að tveggja andlaga sagnir eru ekki óendanlega margar, og leitaði að dæmum um fallhætti eins margra þeirra og ég gat. Með því móti hafði ég fjölmörg dæmi upp úr krafsinu, og sú leit skilaði niðurstöðum sem ég þóttist sjá kerfi í.

En hvort sem ég hef nú rambað á rétta niðurstöðu í þessu tilviki eða ekki (ég hef ekki rekist á neitt síðan sem kollvarpi henni) þá er ljóst að þessi aðferð er ófullnægjandi. Ein ástæða er sú að hún er afskaplega seinleg; ég þurfti að slá inn margar sagnir og leita hvað eftir annað. Önnur ástæða er sú að mikil hætta er á villum; leitin skilar mörgum dæmum sem flest koma ekki málinu við og þegar farið er yfir þau má búast við að eitthvað fari fram hjá manni. Þriðja ástæðan er svo sú að þau orðaraðarmynstur sem ég fann engin dæmi um gæti verið að finna hjá einhverri sögn sem mér datt ekki í hug að leita að.

3.2 Trjábankar

Enn vandast málið ef við ætlum að leita að dæmum um setningagerðir á við kjarnafærslu andlags í aukasetningum, þ.e. dæmum þar sem andlag stendur næst á eftir aukatengingu, eins og í (4):

(4) Ég veit að þennan mann þekkir þú ekki.

Vissulega tengist þessi setningagerð afmörkuðum hópi orða, þ.e. aukatengingum, en vandinn er sá að dæmin um þær eru gífurlega mörg og ef þarf að skoða þau öll jafngildir það því að fara gegnum allan textann. Auðvitað er hægt að þrengja leitina með því að tilgreina tiltekin andlög, en þar með verður leitin líka mjög tilviljanakennd. Þarna dugir orðaleitin því ekki, heldur þyrftum við að geta leitað eftir setningafræðilegu hlutverki – leitað að *andlagi í upphafi aukasetningar*. En til að leita á þann hátt þyrftum við að hafa setningafræðilega greinda texta eða málheild.

Slíkar málheildir eru sums staðar til og mjög víða í smíðum um þessar mundir. Þær eru yfirleitt kallaðar *treebanks*, trjábankar, með vís-

un til þeirrar vel þekktu aðferðar að sýna setningafræðilega formgerð með trjám eða hríslum. Þessir trjábankar eru þó með ýmsu móti og því fer fjarri að í þeim öllum sé formgerð greind á sama hátt og gert er í hríslum. Í sumum trjábönkum (t.d. þeim búlgarska, sjá Osenova og Simov 2003) byggist greiningin á ákveðnu setningafræðilegri kenningakerfi – margir trjábankar sem nú eru í smíðum byggjast t.d. á HPGS, head-driven phrase structure grammar, eða hausastýrðri liðgerðarmálfræði ef við þýðum það á íslensku. Það er sama kenningakerfi og byggt var á í setningagreiningarverkefni því sem unnið var að hjá Friðriki Skúlasyni í nokkur ár (sjá Maren Albertsdóttur og Stefán Einar Stefánsson 2004). Aðrir trjábankar byggjast á venslamálfræði (dependency grammar), t.d. nýr danskur trjábanki (sjá Kromann 2003). Í enn öðrum trjábönkum er áhersla lögð á að hafa greininguna óháða setningafræðilegum teoríum, og þá verður hún yfirleitt ekki jafn smámunasöm (sjá t.d. Nivre 2002).

Ástæðan fyrir því að nú er víða verið að koma upp trjábönkum er sú að úr þeim má vinna mjög margvíslegar upplýsingar um setningagerð – upplýsingar sem ekki fást á annan hátt. Þessar upplýsingar eru ekki síst notaðar í ýmsum verkefnum innan tungutækni, s.s. við málfarsleiðréttingar, vélrænar þýðingar o.fl., en vitanlega nýtast þær einnig við setningafræðirannsóknir, í orðabókagerð o.s.frv. Vandinn er hins vegar sá að setningafræðileg greining samfelldra texta er mjög snúin og tímafrek og stofnkostnaður trjábanka því mjög hár. Reyndar var um tíma í gangi norrænt samstarfsnet um trjábanka, *Nordic Treebank Network*. Í því var gerð tilraun með setningafræðilega greiningu og samanburð á fyrstu köflunum úr *Veröld Soffíu* eftir Jostein Gaarder – Gunnar Hrafn Hrafnbjargarson sá um greiningu á íslenska textanum. En ekki er fyrirsjáanlegt neitt framhald á þessari tilraun hér á landi.

3.3 Setningafræðileg nýting málfræðilegrar mörkunar

3.3.1 Setningafræði í beygingargreiningunni

En þótt setningafræðilega greind íslensk málheild sé ekki til hefur mikill árangur náðst í málfræðilegri (þ.e., einkum beygingarlegri) greiningu íslenskra texta, eins og m.a. kemur fram í grein Sigrúnar Helgadóttur (2007) í þessu hefti. Hér má sjá dæmi um málsgrein sem búið er að marka.

(5) ég fp1en stökk sfg1eþ á aa eftir aþ strætó nkeþ og c veifaði sfg1eþ , , vagnstjórinn nkeng sá sfg3eþ mig fp1eo og c stoppaði sfg3eþ. . ég fp1en tautaði sfg1eþ takk au og c brosti sfg1eþ til ae hans fpkee um ao leið nveo og c ég fp1en lét sfg1eþ miðann nkeog detta sng. .

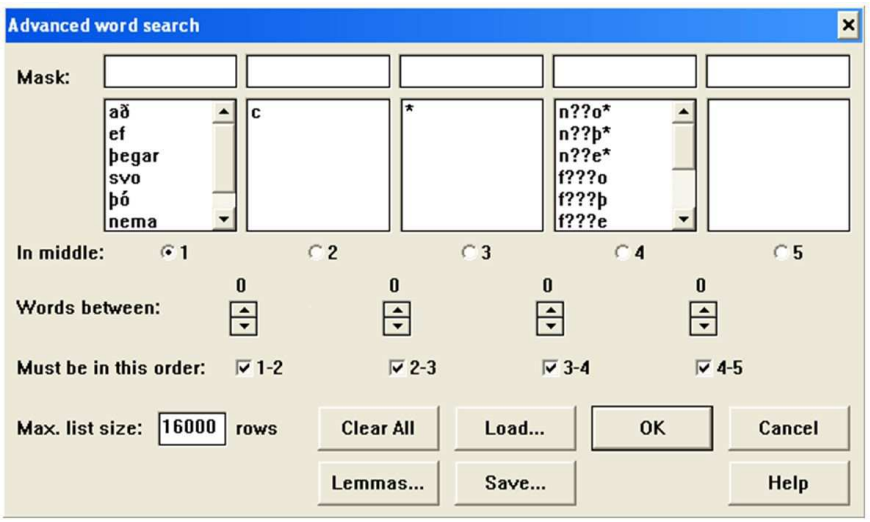
Á eftir hverju orði kemur mark þess eða greiningarstrengur. Hver stafur í strengnum stendur fyrir eitt málfræðilegt atriði. Fyrsti stafurinn stendur alltaf fyrir orðflokk – *f* er fornafn, *s* er sögn, *a* er atviksorð/forsetning, *n* er nafnorð *og* c er samtenging. Aðrir stafir tákna síðan greiningarþætti orðflokkanna. Í *stökk* táknar *s* þannig sögn, *f* framsöguhátt, *g* germynd, *1* 1. persónu, *e* eintölu, og *þ* þátíð. Í *strætó* táknar *n* nafnorð, *k* karlkyn, *e* eintölu og *þ* þágufall. Þótt beygingarlega greiningin taki eingöngu til eiginleika einstakra orða gefa greiningaratriðin mjög oft vísbendingar um vensl orða í setningu; íslensku greiningarstrengirnir gefa miklu meiri setningafræðilegar upplýsingar en þeir ensku t.d. Fallorð innan nafnliðar standa í sama kyni, tölu og falli; frumlag stendur í nefnifalli (nema með skilgreindum hópi sagna) en andlag í aukafalli; o.s.frv. Þess vegna kom sú hugmynd upp að athuga hvort og þá að hvaða marki hægt væri að láta málfræðilegu greiningununa koma í stað setningafræðilegrar greiningar.

Í dæminu hér að framan koma mörkin inn í textann og standa þar eins og hver önnur orð. Það getur oft komið sér vel því að iðulega er hægt að leita að tilteknum setningagerðum með því að tilgreina einhvers konar samband af orðum og greiningarstrengjum.

Ég hef prófað að nota forrit sem heitir *WinCord* til að vinna með þessa mörkuðu texta. Þetta er einfalt forrit sem hægt er að fá ókeypis á netinu, er mjög þægilegt í notkun og hefur gagnast mér vel. En vitanlega væri einnig hægt að nota fjölmörg önnur forrit af svipuðu tagi, eða nota mynsturleit með reglulegum segðum í UNIX eða öðrum tólum. *WinCord* býður upp á samsetta leit (*Advanced Word Search*), þar sem hægt er að slá orð inn í allt að fimm leitarreiti hvern á eftir öðrum, eins og sýnt er hér á eftir. Hér nær 'orð' einnig yfir greiningarstrengi, þar sem þeir koma inn í textann og forritið gerir engan mun á þeim og venjulegum orðum.

3.3.2 Leitað að kjarnafærslu í aukasetningum

Lítum nú á hvernig við förum að því að leita að dæmum um kjarnafærslu í aukasetningum með hjálp WinCord.



Hér höfum við sett sex dæmigerðar aukatengingar inn í fyrsta reitinn, en vitanlega væri hægt að hafa þær fleiri. Í næsta reit er svo mark þessara tenginga, *c*. Þetta er hvort tveggja nauðsynlegt. Ef við notum bara markið en sleppum orðunum fáum við líka allar aðaltengingar með og það viljum við ekki. Ef við notum bara orðin en sleppum markinu fáum við dæmi þar sem orðin í fyrsta reit eru annað en tengingar. Í þriðja reit er stjarna sem stendur fyrir hvað sem er. Það er vegna þess að við viljum ekki þurfa að tilgreina nein tiltekin orð þarna – bara greiningarstreng þeirra sem kemur í fjórða reit. Þar setjum við þrjú mynstur nafnorða, eitt fyrir hvert fall. Spurningarmerki stendur fyrir einn staf, og við höfum tvö spurningarmerki á eftir *n* vegna þess að hvorki kyn né tala skipta máli. Aftast setjum við svo stjörnu sem stendur fyrir ótiltekinn fjölda stafa. Það er vegna þess að ekki skiptir máli hvort orðið hefur viðskeyttan greini, er mannsnafn eða staðarnafn, o.s.frv. Þarna eru líka þrjú mynstur fornafna, þar sem undirflokkur, kyn og tala skiptir ekki máli; og einnig eru þarna (þótt það sjáist ekki á myndinni) þrjú mynstur lýsingarorða, eitt fyrir hvert aukafall.

Leit með þessu mynstri í textum *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) skilar á fimmta hundrað dæmum – hér má sjá nokkur þeirra.

(6) a. ég held að henni leiðist vinnan
b. og af hverju ekki ef mér leyfist að spyrja?
c. það var óneitanlega léttir þegar þessu var lokið
d. allt og sumt var að hana hafði dreymt eitthvað um nóttina

Eins og þessi dæmi benda til sýna langflest þeirra dæma sem leitin skilar í raun aukafallsfrumlög – skilgreining okkar dugir ekki til að greina þar á milli. Með því að láta forritið raða dæmunum eftir sögninni er þó mjög fljótlegt að vinsa þessi dæmi frá. Í sumum þeirra dæma sem þá standa eftir er aukafallsnafnliðurinn ekki andlag heldur hefur stöðu atviksliðar, eins og hér eru sýnd dæmi um:

(7) a. og vill svo til að þann dag varð Tryggvi Gunnarsson sjötugur
b. og um leið og þetta upplaukst fyrir henni varð henni ljóst að allar götur frá fyrstu árum hafði hún verið að lemja niður einhvern óskiljanlegan ótta

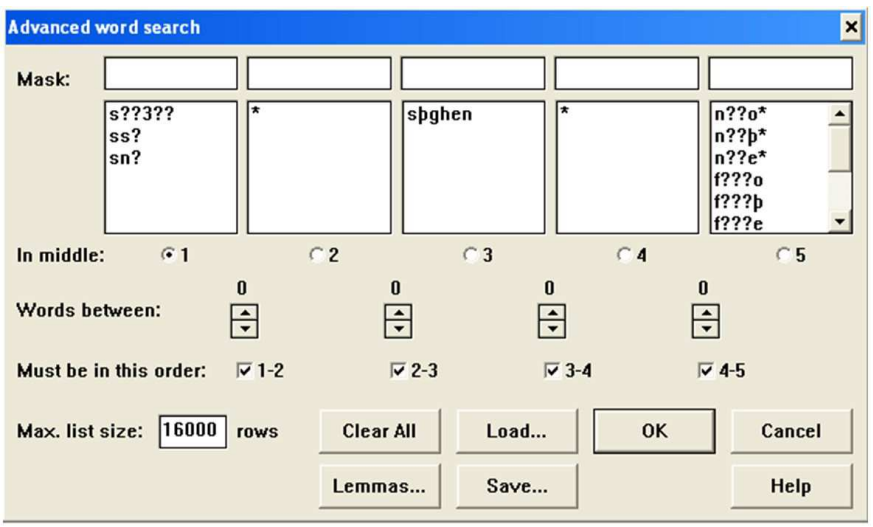
Þegar þessi dæmi hafa líka verið síuð frá koma í ljós fáein dæmi um kjarnafærslu, eins og hér er sýnt.

(8) a. Guð veit að það geri ég líka
b. hún fullyrti að það myndi hún gera
c. ég gerði hins vegar alls ekki ráð fyrir að þessu sæti myndi fylgja seta á Alþingi
d. þeir vissu þó vel að þetta máttu þeir ekki
e. reynt var að leiða henni fyrir sjónir að þetta yrðu allir að gera
f. Vigdís forseti segir að bestu ljóð um vor og sumar hafi karlar ort til kvenna

Þarna er sem sé hægt á fáeinum mínútum að kalla fram öll dæmi um kjarnafærslu andlaga í aukasetningum í 500 þúsund orða texta. Það hefði tekið fleiri daga að finna þessi dæmi í ómörkuðum texta, hvort sem maður hefði leitað að þeim með því hreinlega að lesa allan textann eða reynt að leita að dæmunum í tölvu á einhvern hátt.

3.3.3 Leitað að nýju þolmyndinni

Annað dæmi má taka af hinni svokölluðu „nýju þolmynd“ (Sigríður Sigurjónsdóttir og Joan Maling 2001). Megineinkenni hennar er að andlag germyndarsetningar flyst ekki í frumlagssæti þótt sögnin taki á sig þolmyndarform (þ.e., aðalsögnin standi í lýsingarhætti þátíðar og so. *vera* eða *verða* komi inn), heldur stendur áfram í dæmigerðu andlagssæti. Þar að auki halda þolfallsandlög falli sínu í stað þess að fá nefnifall eins og þau gera þegar þau færast í frumlagssæti í venjulegri þolmynd (þágufalls- og eignarfallsandlög halda aftur á móti alltaf falli sínu). Að þessari setningagerð má leita með þessu mynstri:



Hér táknar * í öðrum leitarreit og *spghen* í þriðja leitarreit að leitað sé að lýsingarhætti þátíðar af hvaða sögn sem er. Þar á eftir kemur óskilgreint orð, en í fimmta leitarreit er tilgreint að það orð skuli vera nafnorð eða fornafn (eða lýsingarorð, þótt það sjáist ekki á myndinni) í aukafalli. Í fyrsta reit getur svo verið sögn í þriðju persónu (*s??3??*), nafnhætti (*sng*) eða sagnbót (*ssg*). Þannig finnast dæmi eins og *(Það) var barið mig*, *(Það hefur) verið barið mig*, og *(Það mun) verða barið mig*.

Þetta leitarmynstur var keyrt á texta *Íslenskrar orðtíðnibókar* (Jörgen Pind, Friðrik Magnússon og Stefán Briem 1991) og skilaði 45 dæmum, en fljótlegt er að ganga úr skugga um að ekkert þeirra sýnir nýja þolmynd svo að öruggt sé. Þarna eru t.d. 8 dæmi um sambandið *þegar/er hér var komið sögu*. Eina dæmið sem svipar til nýrrar þolmyndar

er eftirfarandi:

(9) þar með var lokið hvellinum mikla

Lítill vafi leikur þó á því að eðlilegra er að greina þetta dæmi á annan hátt (þ.e. sem frestun þungs nafnliðar eða eitthvað slíkt; sjá Höskuld Þráinsson 2005:587–588).

Sama leitarmynstur var keyrt á ÍSTAL-safnið og skilaði þar fimm dæmum; eitt þeirra gæti hugsanlega verið ný þolmynd:

(10) það var lokað tjaldstæðinu á Þingvöllum

Hér er þó einnig hugsanlegt að um sé að ræða frestun þungs nafnliðar en ekki nýja þolmynd. Til að fá ótvíræð dæmi um nýja þolmynd þyrfti andlagið að vera perónufornafn, því að þeim er ekki hægt að fresta á þennan hátt.

Það kemur ekki á óvart að engin ótvíræð dæmi um nýja þolmynd skuli finnast í þessum tveimur söfnum. Bæði er þessi setningagerð yfirleitt talin frekar nýtilkomin, og þar að auki að mestu bundin við mál barna og unglinga (sjá Sigríði Sigurjónsdóttur og Joan Maling 2001), en allir textar í söfnunum eru frá fullorðnu fólki. En vissulega er gagnlegt að geta fengið staðfestingu á þessu með ítarlegri leit.

3.3.4 Leitað að *það*-lepp með áhrifssögnum

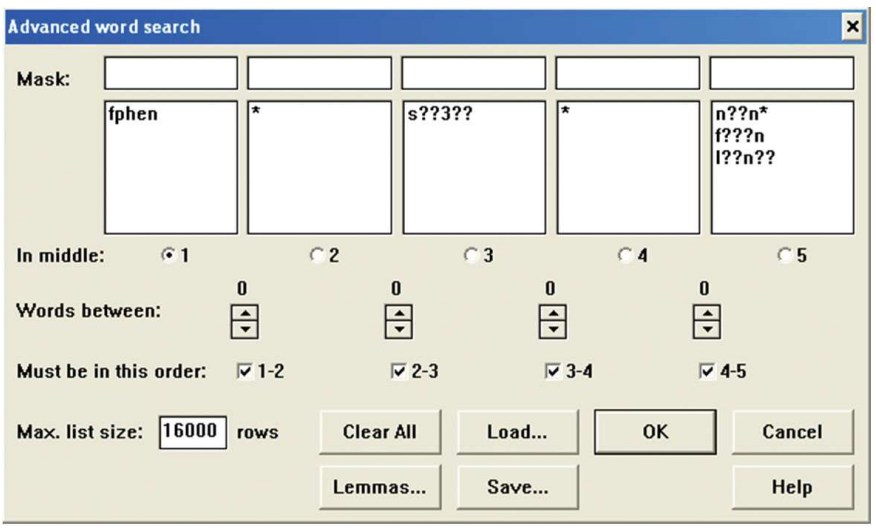
Dæmi má einnig taka af leppnum eða aukafrumlaginu *það* (sjá Eiríks Rögnvaldsson 2002). Ég leitaði einu sinni í ÍSTAL-safninu að dæmum um *það* með áhrifssögnum, eins og sést í (11), og sagðist „í fljótu bragði“ ekki hafa fundið nein dæmi af þessu tagi þar (Eiríkur Rögnvaldsson 2002:11nm):

(11) a. Það hefur einhver borðað allan grautinn minn.
b. Það getur enginn svarað þessu.
c. Það stungu einhverjir stúdentar smjörinu í vasann.
d. Það keypti hann eitthvert fífl.

Á þeim tíma átti ég ekki völ á öðru en textaleit; leitaði að dæmum um *það* og fór yfir þau. En þau dæmi eru ákaflega mörg í ÍSTAL (á sjöunda þúsund) þannig að auðvelt er að láta sér sjást yfir einhver dæmi um það sem maður er að svipast um eftir. Vegna þess hversu

seinlegt þetta var lagði ég ekki í að leita í öllu ÍSTAL-safninu í þessari leit.

Mér fannst þess vegna forvitnilegt að gera aðra atrennu að því að leita að þessari setningagerð í ÍSTAL, og notaði til þess eftirfarandi mynstur:



Hér er *fphen* í fyrsta reit; það á eingöngu við *það*. Annar og þriðji reitur skilgreina svo ótilgreina sögn í 3. persónu, og fjórði og fimmti reitur skilgreina fallorð (nafnorð, fornafn eða lýsingarorð) í nefnifalli. Þessi leit skilar 720 dæmum úr ÍSTAL, en í þeim flestum er nefnifallsliðurinn sagnfylling en ekki frumlag. Hann stendur þá næst á eftir einhverri mynd so. *vera* eða *verða*, og þeim dæmum er auðvelt að henda út með því að láta forritið raða dæmunum eftir sögninni. Þá standa eftir milli 30 og 40 dæmi, sem sum hver sýna ótvírætt þá setningagerð sem leitað var að:

(12) a. það átti enginn skap saman
b. það þekkja allir Rósu
c. það vita allir hver Rósa er
d. það heldur enginn að þú sért hommi

Í stað þess að lesa allt ÍSTAL-safnið, eða fara gegnum á sjöunda þúsund dæma um *það*, dugir því að skoða þessi 30-40 dæmi. Í stað margra tíma þreytandi yfirlegu þar sem hætta á mistökum er veruleg

kemur 10-20 mínútna vinna þar sem tækifæri gefst til að skoða hvert dæmi vandlega og meta hvort það falli undir þá setningagerð sem leitað er að.

3.4 Setningafræðileg þáttun

Þessar tilraunir sýna ótvírætt að hægt er að hafa verulegt gagn af hinni beygingarlegu mörkun í setningafræðilegri dæmaleit. Vissulega er ekki hægt að leita að öllum setningagerðum á þennan hátt. Það á m.a. við um setningagerðir þar sem vensl milli orða ná yfir ótiltekinn orðafjölda, langdræg vensl. Það væri t.d. erfitt að nota þessa aðferð til að skoða vísun afturbeygðra fornafna svo að dæmi sé tekið. *WinCord* forritið býður að vísu upp á að tilgreint sé hversu mörg orð megi koma á milli orðanna í leitarreitunum – jafnvel er hægt að leyfa ótiltekinn orðafjölda. En gallinn við það er að þá kemur óhjákvæmilega með mikill fjöldi dæma sem ekki koma málinu við, og mjög tímafrekt getur verið að hreinsa frá.

Nú er að opnast annar möguleiki á setningafræðilegri leit í íslenskum textum. Verið er að vinna að gerð hlutaþáttara (e. shallow parser) fyrir íslensku. Hlutaþáttun er setningafræðileg greining þar sem ekki er stefnt að því að sýna fullkomna formgerð setninga eða öll vensl liða. Þess í stað er lögð áhersla á að greina helstu setningarliði – flokka saman orð sem eiga saman. Slík greining getur nýst vel í ýmsum tilvikum, og hentar stundum betur en full greining (e. deep parsing). Einnig eru helstu setningafræðileg hlutverk greind. Samið hefur verið sérstakt þáttunarskema til að skilgreina hvaða liðir og hlutverk eru greind, og við hvað skuli miðað í greiningunni (Hrafn Loftsson og Eiríkur Rögnvaldsson 2006). Hér er sýnt dæmi um setningu sem greind hefur verið eftir þessu skema.

(13) {***SUBJ**> [*NP* augnaráðið `nheng` *NP*] ***SUBJ**>}
[*VP* negldist `sfm3eþ` *VP*]
[*PP* við `ao` [*NP* [*AP* gráa `lkeovf` AP] jakkann `nkeog` *NP*] *PP*]
[*SCP* sem `ct` *SCP*]
{***SUBJ**> [*NP* hann `fpken` *NP*] ***SUBJ**>}
[*VPb* var `sfg3eþ` *VPb*]
[*VPi* að `cn` klæða `sng` *VPi*]
{***OBJ**< [*NP* sig `fpkeo` *NP*] ***OBJ**<}
[*PP* úr `aþ` *PP*]
[*CP* og `c` *CP*]
[*VPi* hengja `sng` *VPi*]
[*PP* [*MWE_PP* inn `aa` í `ao` *MWE_PP*] [*NP* skáp `nkeo` *NP*] *PP*]

Hér eru beygingarleg mörk með lágstöfum; mörk setningarliða með hástöfum og skáletruð, auk þess sem hornklofar umlykja liðina; og mörk setningafræðilegra hlutverka hefjast á stjörnu og eru með hástöfum og feitletruð, og slaufusvigar afmarka hlutverkin. Mörk liða og hlutverka ættu flest að vera auðskilin. Þó er rétt að nefna að *MWE* stendur fyrir 'multiword expression' og er notað til að marka orðarunur sem í raun eru ígildi eins orðs (einkum fleiryrtar samtengingar og forsetningar). Oddur (> og <) er notaður á frumlög og andlög (og sagnfyllingar) til að vísa á sögnina sem þessir liðir tengjast.

Eins og sjá má er hér ekki gerð tilraun til að tengja liði saman og sýna þannig stigveldisformgerð setningarinnar, nema að litlu leyti (nafnliðir eru sýndir sem hluti forsetningarliða). Áhersla er fremur lögð á að ná mikilli nákvæmni í vélrænu greiningunni, miðað við greiningarskemað sem þáttarinn vinnur með. Fyrstu tilraunir benda til að þáttarinn skili hlutverki sínu mjög vel og villur í greiningunni séu fáar.

Þar með eru komnar nýjar forsendur til setningafræðilegrar dæmaleitar í textum. Með þessari greiningu verður t.d. hægt að leita að röðinni *aukatenging – andlag – sagnliður* og finna þannig dæmi um kjarnafærslu í aukasetningum, í stað þess að byggja leitina á beygingarlegum mörkum eins og gert var hér að framan. Það á svo eftir að koma í ljós hvort leit eftir setningafræðilegum mörkum skilar betri árangri en hin. En hér eru e.t.v. að opnast möguleikar á að koma upp vísi að íslenskum trjábanka.

4 Lokaorð

Í fyrri hluta þessarar greinar var sagt lauslega frá mismunandi viðhorfum til gildis textadæma í setningafræði undanfarna áratugi. Tilkoma generatífrar málfræði fyrir hálfri öld olli því að dæmaleit í textum naut lítillar virðingar um langt skeið, og málfræðingar skiptust í fylkingar sem höfðu mjög andstæðar skoðanir á þessu sviði, þótt sá ágreiningur hafi að verulegu leyti verið sýndarágreiningur og stafað af því að menn voru að bera saman epli og appelsínur. En með tilkomu viðamikilla rafrænna textasafna og málheilda, og ekki síst mikils magns veftexta, hafa skilin milli fylkinganna dofnað og nú þykir ekki lengur neitt að því að safna dæmum úr textum. En ýmislegt er að varast við notkun dæmanna og gæta verður varúðar í túlkun þeirra, eins og bent er á í greininni.

Meginviðfangsefni greinarinnar var að skoða hvernig hægt er að standa að verki við leit að dæmum um tilteknar setningagerðir í rafrænum íslenskum textasöfnum. Bent var á að hægt er að nýta hráa texta, án nokkurrar sérstakrar mörkunar, að vissu marki, en þó því aðeins að leitað sé að setningagerðum sem tengjast ákveðnum orðum. Með tilkomu beygingarlega markaðra málheilda og mörkunarforrita hafa aðstæður til setningafræðilegrar leitar hins vegar gerbreyst. Vegna eðlis íslenska beygingakerfisins má lesa miklar setningafræðilegar upplýsingar út úr hinum beygingarlegu mörkum, og þær upplýsingar má síðan nýta í leit að ákveðnum setningagerðum. Sýnd voru þrjú dæmi um hvernig hægt er á einfaldan og fljótvirkan hátt að leita að dæmum um þrjár setningagerðir; kjarnafærslu í aukasetningum, nýja þolmynd, og *það*-lepp með áhrifssögnum. Í öllum tilvikum skilaði leitin niðurstöðum sem tekið hefði fleiri daga að fá með þeim aðferðum sem áður var völ á, en nú tók leitin aðeins fáeinar mínútur.

Í lok greinarinnar er svo sagt frá verkefni sem enn er ólokið og felst í gerð hlutaþáttara fyrir íslensku. Ef það verkefni skilar tilætluðum árangri er hægt að fara að gera raunhæfar áætlanir um smíði viðamikils íslensks trjábanka sem myndi verða öllum sem fást við rannsóknir á íslenskri setningafræði að ómetanlegu gagni.

Heimildir

Ásta Svavarsdóttir. 2006. Tilbrigði í setningagerð. *Orð og tunga* 8:156–157.

Ásta Svavarsdóttir. 2007. Talmál og málheildir – talmál og orðabækur. *Orð og tunga* 9 (þetta hefti).

Bergljót S. Kristjánsdóttir, Eiríkur Rögnvaldsson, Guðrún Ingólfsdóttir og Örnólfur Thorsson (ritstj.). 1996. *Orðstöðulykill Íslendinga sagna.* [Geisladiskur.] Mál og menning, Reykjavík.

Chomsky, Noam. 1957. *Syntactic Structures.* Mouton, Haag.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Massachusetts.

Eiríkur Rögnvaldsson. 1994-95. Breytileg orðaröð í sagnlið. *Íslenskt mál* 16–17:27–66.

Eiríkur Rögnvaldsson. 1996. Frumlag og fall að fornu. *Íslenskt mál* 18: 37–69.

Eiríkur Rögnvaldsson. 1998. Heimildatúlkun í sögulegri setningafræði. Baldur Sigurðsson, Sigurður Konráðsson og Örnólfur Thorsson (ritstj.): *Greinar af sama meiði*, bls. 317–334. Rannsóknarstofnun Kennaraháskóla Íslands, Reykjavík.

Eiríkur Rögnvaldsson. 2000. Setningarstaða boðháttarsagna í fornu máli. *Íslenskt mál* 22:63–90.

Eiríkur Rögnvaldsson. 2002. *ÞAÐ* í fornu máli – og síðar. *Íslenskt mál* 24:7–30.

Faarlund, Jan Terje. 1990. *Syntactic Change. Toward a Theory of Historical Syntax.* Mouton, Berlín.

Falk, Cecilia. 1995. Lexikalt kasus i svenska. *Arkiv för nordisk filologi* 110:199–226.

Halldór Ármann Sigurðsson. 1989. *Verbal Syntax and Case in Icelandic.* In a Comparative GB Framework. Doktorsritgerð, Lund Universitet, Lund.

Hrafn Loftsson og Eiríkur Rögnvaldsson. 2006. A Shallow Syntactic Annotation Scheme for Icelandic Text. *Technical Report RUTR-SSE06004*, Department of Computer Science, Reykjavik University, Reykjavík.

Höskuldur Þráinsson. 1979. *Complementation in Icelandic.* Garland, New York.

Höskuldur Þráinsson. 2005. *Setningar.* Handbók um setningafræði. (Íslensk tunga III.) Almenna bókafélagið, Reykjavík.

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

Kromann, Matthias Trautner. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. Joakim Nivre og Erhard Hinrichs (ritstj.): *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, bls. 217–220. Växjö University Press, Växjö.

Lightfoot, David S. 1979. *Principles of Diachronic Syntax.* Cambridge University Press, Cambridge.

Maren Albertsdóttir og Stefán Einar Stefánsson. 2004. Beygingar- og málfræðigreinikerfi. *Samspil tungu og tækni*, bls. 16–19. Menntamálaráðuneytið, Reykjavík.

McEnery, Tony, og Andrew Wilson. 1996. *Corpus Linguistics.* Edinburgh University Press, Edinburgh.

Mørck, Endre. 1992. Subjektets kasus i norrønt og mellomnorsk. *Arkiv för nordisk filologi* 107:53–99.

Nivre, Joakim. 2002. What kinds of trees grow in Swedish soil? A comparison of four annotation schemes for Swedish. Erhard Hinrichs og Kiril Simov (ritstj.): *Proceed-*

ings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002), 20–21 September 2002. Sozopol, Bulgaria.
Osenova, Petya, og Kiril Simov. 2003. The Bulgarian HPSG Treebank: Specialization of the Annotation Scheme. Joakim Nivre og Erhard Hinrichs (ritstj.): *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, bls. 129–140. Växjö University Press, Växjö.
Sigríður Sigurjónsdóttir og Joan Maling. 2001. Það var hrint mér á leiðinni í skólann: Þolmynd eða ekki þolmynd? *Íslenskt mál* 23:123–180.
Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga* 9 (þetta hefti).
Þórunn Blöndal. 2005. *Lifandi mál.* Inngangur að orðræðu- og samtalsgreiningu. Rannsóknarstofnun Kennaraháskóla Íslands, Reykjavík.

Lykilorð:

málheildir, dæmasetningar, málfræðileg mörkun

Keywords:

text corpora, example sentences, PoS tagging

Abstract

This paper discusses the use of text corpora in syntactic research, and how to search for example sentences in corpora. During the past few decades, widely divergent views have been expressed as to the value of corpora in syntactic argumentation. It is argued in the paper that this disagreement stems from different views as to the subject of linguistic research. The paper also discusses various problems that arise in the interpretation of the information extracted from corpora – especially in drawing conclusions from the silence of the texts on certain constructions. The main section of the paper discusses the possibilities of searching for certain syntactic constructions in different types of Icelandic corpora; raw untagged text, PoS tagged text, and text where the major syntactic constituents and syntactic functions have been identified. Data-driven PoS taggers have now been trained on Icelandic texts, and it is shown that due to the inflectional character of Icelandic and the richness of the tagset, the resulting PoS tagging is very effective in the search for various syntactic constructions.

Eiríkur Rögnvaldsson
Háskóla Íslands
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is