

Eiríkur Rögnvaldsson

# Setningafræði í tungutækni

## 1. Inngangur

Á undanförunum árum hefur þróunin í margs konar samskiptum manns og tölvu verið ör, ekki síst á sviði **tungutækni** (e. language technology, sjá Eirík Rögnvaldsson 2001). Sá tími er á næstu grösom að við viljum geta talað við tölvurnar okkar og ætlumst til að þær skilji okkur og geti svarað „af viti“, rétt eins og maður talar við mann. Slík samskipti eiga sér nú þegar stað sums staðar erlendis, t.d. í ýmiss konar upplýsingakerfum þar sem menn leita sér fróðleiks um afmörkuð efni eins og flugferðir, lestarsamgöngur, veitingastaði, leikhús o.s.frv. Þess konar kerfi eru víða í þróun og misjafnlega á vegi stödd. En margt þarf að koma til ef þau eiga að geta þjónað tilgangi sínum – upptalningin hér á eftir er mjög einfölduð:

Kerfið þarf að

- greina segðina (fyrirspurnina) hljóðfræðilega
- varpa greiningarstrengnum yfir í orð sem það þekkir
- greina setningafræðilega formgerð segðarinnar
- greina merkingu segðarinnar (út frá orðum og formgerð)
- leita upplýsinga í gagnasafni sínu til að mynda svar
- beita setningafræðilegum reglum til að mynda andsvar við segðinni
- breyta setningunni í hljóð

Ef einn þessara þátta bregst fara samskiptin út um þúfur og viðmælandi kerfisins fær ekkert vitrænt svar við fyrirspurn sinni. Greini kerfið setningagerðina ekki rétt er t.d. hætt við að merkingartúlkunin misfarist.

En vélræn **þáttun** (e. parsing) setninga er gagnleg og nauðsynleg miklu víðar en í upplýsingakerfum. Hún er t.d. forsenda þess að hægt sé að smíða hugbúnað til að leiðrétta texta málfræðilega (sjá Eirík Rögnvaldsson 2001). Aðeins lítt hluti málfarsvillna felst í því að notaðar séu orðmyndir sem ekki eiga að koma fyrir í málinu (t.d. *föðurs* í stað *föður*, *keyptu* í stað *kauptu*). Langflestar villur felast í því að nota leyfilegar orðmyndir á óleyfilegum stöðum í setningu. Villur eins og *Ég hitti systir þína* (í stað *systur*), *vegna þeirrar tilhneigingu* (í stað *tilhneigingar*), *fjöldi manna komu* (í stað *kom*), *mér langar* (í stað *mig langar*) o.s.frv. er ekki hægt að finna og leiðrétta á vélrænan hátt án málfræðilegrar greiningar og setningafræðilegrar þáttunar, því að *systir*, *tilhneigingu*, *mér* og *komu* eru allt fullkomlega leyfilegar íslenskar orðmyndir – bara ekki á þessum stöðum í setningu.

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

Setningafræðileg þáttun er einnig mikilvægur þáttur í vélrænum þýðingum. Allir vita að til að þýða texta af einu máli á annað er ekki nóg að fletta upp í orðabók

og finna þýðingu hvers orðs fyrir sig. Setningagerð tungumála er misjöfn og þýðingin verður að taka tillit til þess – breyta orðaröð þar sem það á við o.s.frv. Hér má sjá tvö dæmi um vélræna þýðingu úr ensku á íslensku (frá þýðingarforriti sem Stefán Briem hefur í smíðum, sjá <http://www.simnet.is/stbr/stada.html>):

- (1) a. Since then, Reykjavik has grown exponentially.  
b. Síðan þá, Reykjavík hefur vaxið með veldishraða.
- (2) a. Do you ever wonder what is out in space?  
b. Gera þig ávallt furðar hvað er út í rúmi?

Hér er ekkert orð beinlínis ranglega þýtt (þótt vissulega væri heppilegra að þýða *space* með *geimur* en *rúm* í (2b)). En í fyrra dæminu tekur þýðingarforritið ekki tillit til þess að í íslensku er sögnin ekki aftar en í öðru sæti setningar – *hefur* þyrfti að koma á undan *Reykjavík*. Í seinna dæminu veit forritið ekki að íslenskar spurningar eru myndaðar án sérstakrar hjálparsagnar, ólíkt því sem er í ensku.

## 2. Vélræn þáttun og setningafræðileg líkön

Ýmis líkön og kenningakerfi hafa verið notuð við vélræna þáttun setninga. Menn hafa mikið rökrætt hvort þau líkön og fræðikenningar sem notuð eru í fræðilegum málvísindum gagnist eða eigi við í vélrænni þáttun. Flestir eru þó á þeirri skoðun að setningarlýsing af þeirri gerð sem byggt er á í þessari bók sé ekki sérlega heppileg í vélrænni þáttun. Ástæðan er einkum sú að þar er einatt gert ráð fyrir einhverjum mun – oft verulegum – á grunngerð setninga og þeirri gerð sem við heyrum eða sjáum. Þessi munur stafar ekki síst af ýmiss konar færslum, s.s. sagnarfærslu, nafnliðafærslum, kjarnafærslu o.s.frv. Greiningarforrit hafa vitaskuld aðeins aðgang að endanlegri gerð setninganna, og í þeirri gerð er oft erfitt að finna vísbendingar um að tiltekin færsla hafi átt sér stað. Þess vegna er mjög erfitt fyrir greiningarforrit að greina grunngerð út frá endanlegri gerð.

En þótt líkan eins og það sem hér er notað henti ekki í vélrænni greiningu þarf það ekki að draga úr málfræðilegu gildi þess. Í fyrsta lagi eru tölvur ekki mannlegar, og ekki sjálfgefið að mönnum takist nokkurn tíma að koma upp málgreiningarhugbúnaði sem verði sambærilegur við mannshugann. Það á a.m.k. langt í land. Þar af leiðandi er ekki heldur sjálfgefið að greiningarlíkön sem notuð eru við lýsingu á málkunnáttu manna komi að gagni þegar verið er að kenna tölvum málgreiningu. Í öðru lagi lýsa fræðileg mállíkön málkunnáttunni frá sjónarmiði talandans. Þau gera grein fyrir því hvernig endanleg gerð setningar er leidd út frá grunngerðinni. Greiningarforrit hafa hins vegar aðeins aðgang að endanlegri gerð, eins á áður er bent á, og verða að greina formgerðina út frá henni. Viðfangsefnið er því í raun allt annað en í fræðilegri setningafræði og engin ástæða til að gera ráð fyrir að sams konar líkön eigi endilega við.

Sum setningafræðileg kenningakerfi eru þó bæði notuð við lýsingu á málkunnáttu manna og í smíði greiningarhugbúnaðar fyrir tölvur. Þar má einkum nefna tvö vel þekkt kerfi: **LFG** – Lexical-Functional Grammar (Orðstýrða hlutverkamálfræði, Kaplan og Bresnan 1982) og **HPSG** – Head-Driven Phrase Structure Grammar (Höfuðstýrða liðgerðarmálfræði, Pollard og Sag 1994). Þessi kerfi nota bæði **samhengisfrjálsar liðgerðarreglur** (e. context-free phrase structure rules), þ.e. reglur á við  $S \rightarrow Nl Sl$ , til að lýsa gerð setninga. Þau eiga það sameiginlegt að gera ekki ráð

fyrir afleiðslu frá einni gerð (grunngerð) til annarrar, og vinna því hvorki með færslur né spor, en sýna sambærilegar upplýsingar um tengsl orða og liða á annan hátt, s.s. með **sammerkingu** (e. co-indexing), **valhömlum** (e. selectional restrictions) o.fl.

Bæði þessi kerfi eru upprunnin innan fræðilegra málvísinda fyrir u.þ.b. aldarfjórðungi, en hafa lengi verið notuð innan tölvumálvísinda og hafa því lagað sig að þeim á ýmsan hátt. Sem dæmi um nýtingu þeirra á því sviði má nefna stóra norska þýðingarverkefnið LOGON, sem miðar að vélrænum þýðingum úr norsku á ensku og á að ljúka 2007 (Lønning o.fl. 2004). Þar er norski frumtextinn greindur með hugbúnaði sem byggist á LFG, en HPSG er notað til að mynda setningar ensku þýðingarinnar. Tilraunir með kerfið benda til að þessi aðferðafræði gangi upp, þótt fullkomin vélræn þýðing eigi enn langt í land.

Eina stóra verkefnið í vélrænni þáttun íslenskra setninga sem unnið hefur verið að byggðist einnig á HPSG, en Friðrik Skúlason stóð fyrir því á árunum 2002-2005 (sjá Maren Albertsdóttur og Stefán Einar Stefánsson 2004). Í því verkefni var unnið mikið starf við að skilgreina íslenskt málkerfi í HPSG, en að öðru leyti hefur fremur lítið verið unnið með íslensku innan LFG og HPSG.

### 3. Aðferðir við vélræna þáttun

Vélræn þáttun setningar felst í því að fara í gegnum alla hugsanlega greiningarmöguleika, eða **þáttunartré** (e. parse tree), þeirrar setningar sem er til skoðunar og reyna að finna það (þau) tré sem á (eiga) við. Þessi leit stýrist af tvennu; annars vegar inntaki greiningarforritsins, þ.e. orðunum í setningunni, og hins vegar þeirri málfræði sem forritið vinnur út frá. Forritið reynir að finna (eða búa til) tré með rótina  $S$  (vegna þess að strengurinn er setning) þar sem orðin í strengnum tengjast þessu  $S$  í samræmi við þær liðgerðarreglur sem forritið kann. Þessi leit er fjarri því að vera einföld, og alls ekki gefið að takist að þátta allar setningar. Bæði eru reglurnar sem þarf að prófa margar og gefa aragrúa möguleika þegar þeim er beitt aftur og aftur; og eins er fjölbreytileikinn í setningum málsins ærinn. En við leitina er hægt að beita tveimur mismunandi aðferðum (sjá Jurafsky og Martin 2000).

Annars vegar getur leitin verið **ofansækin** (e. top-down). Þá byrjar hún á  $S$  efst í trénu og býr til allar hugsanlega trjábúta sem byrja á  $S$  út frá öllum liðgerðarreglum sem hafa  $S$  vinstra megin við ör ( $S \rightarrow Nl Sl, S \rightarrow Sl, \text{o.s.frv.}$ ). Síðan er haldið áfram með þessa trjábúta í næsta lagi, tekin öll ták næst neðan við  $S$  ( $Nl, Sl \text{ o.s.frv.}$ ), fundnar reglur þar sem þau eru vinstra megin við ör ( $Nl \rightarrow no, Nl \rightarrow lo no, Sl \rightarrow so, Sl \rightarrow so Nl \text{ o.s.frv.}$ ) og trén teiknuð áfram í samræmi við þær. Þannig er haldið áfram þar til kemur að orðunum í strengnum. Þá er skoðað hvort hægt sé að hengja þau á leyfilegan hátt í einhver þeirra trjáa sem búin hafa verið til. Þeim trjám sem ekki passa við strenginn er hent, og eftir stendur (vonandi) eitt tré (eða hugsanlega fleiri, ef setningin er margræð) sem lýsir setningafræðilegri formgerð strengsins.

Hins vegar getur leitin verið **neðansækin** (e. bottom-up). Þá er byrjað á orðunum í strengnum sem greina skal, þeim flett upp í orðasafni og hugsanleg málfræðileg greining skrifuð út. Síðan er reynt að tengja orðin saman; leitað að strengjum sem passa við það sem er hægra megin við örina í einhverri reglu. Ef strengurinn byrjar t.d. á orðunum *Þessi bók* finnst væntanlega í orðasafni að *þessi* sé ábendingarfornafn og *bók* nafnorð. Þá er athugað hvort röðin *áfn – no* finnist einhvers staðar vinstra megin við ör. Í reglusafninu er væntanlega regla í líkingu við  $Nl \rightarrow áfn no$ , og því er gískað á að *Þessi bók* sé nafnliður. Þannig er haldið áfram að tengja stærri og stærri einingar þar til kemur að  $S$ , og trjám (eða trjábútum) sem ekki leiða þangað er hent.

Báðar aðferðirnar hafa sína kosti og galla. Ofansækin þáttun gengur út frá  $S$  og eyðir því aldrei tíma í trjábúta sem ekki geta endað í  $S$ . Hún horfir ekki á orðin í setningunni fyrr en að þeim kemur og býr því til fjölda trjáa sem ekki falla að gögnunum. Neðansækin þáttun gengur út frá orðunum og eyðir því ekki tímanum í tré sem ekki falla að gögnum. Hún horfir hins vegar ekki á rótina,  $S$ , og býr því til trjábúta sem aldrei geta orðið að heilu tré.

Á seinustu árum hefur hlutverk **tölfræðilegra líkana** (e. statistical models) í setningafræðilegri þáttun farið ört stækkandi samfara þróun í tölvutækni (sjá Manning og Schütze 1999). Til að þjálfa slík líkön þarf tölvutæka og setningafræðilega greinda texta sem fer sífellt fjölgandi, og mikil aukning verður árlega á vinnslugetu, minni og diskrymi tölva, en allt er þetta nauðsynlegt til að hægt sé að vinna með mikið textamagn á stuttum tíma. Hægt er að beita **líkindaþáttun** (e. probabilistic parsing) byggðri á tölfræðilegum líkönum einni sér, en einnig er hægt að láta tölfræðilíkin vinna með hinum málfræðilegu á ýmsan hátt. Þannig væri t.d. hægt að láta HPSG finna út allar hugsanlegar greiningar á tiltekinni setningu, og beita síðan tölfræðilegum upplýsingum til að velja líklegustu greininguna úr hópi þeirra mögulegu.

## 4. Mismunandi ítarleg þáttun

Markmiðið með vélrænni greiningu setninga er oftast hagnýtt, og greiningin miðast því oft við tiltekin verkefni og sérstakar þarfir þeirra. Hún getur þess vegna verið með ýmsu móti eftir tilgangi sínum hverju sinni. Ef á að nota þáttunina í vélrænni þýðingu yfir á annað tungumál veitir líklega ekki af ítarlegri greiningu eins og lýst er í 3. kafla hér að framan, þar sem stigveldisformgerð setninga er greind nákvæmlega og grein gerð fyrir öllum innbyrðis venlum orða og liða. Slík greining er kölluð **djúp þáttun** (e. deep parsing) og er t.d. beitt þegar notuð eru málfræðilíkon eins og LFG og HPSG sem nefnd eru hér að framan.

En ekki er alltaf þörf á nákvæmri greiningu – oft getur verið kappnóg að vita að tiltekin orðaruna er nafnliður sem er frumlag setningarinnar, og önnur orðaruna er nafnliður sem gegnir hlutverki andlags, en engin þörf á að lýsa formgerð setningarinnar í margra hæða stigveldi. Það er kallað **grunn þáttun** (e. shallow parsing). Hún getur dugað vel til að greina meginatriði í merkingu setningar þannig að hægt sé að bregðast við henni á viðeigandi hátt, t.d. í leitar- eða upplýsingakerfi. Við grunna þáttun setninga er oft notuð greiningaraðferð sem nefnist **hömlumálfræði** (e. Constraint Grammar, Karlsson o.fl. 1995). Þar er hverju orði gefið bæði málfræðilegt og setningafræðilegt mark, en setningafræðilegu mörkin eru byggð á **hæði** (e. dependency) milli orða. Í upphafi getur hvert orð fengið mörg setningafræðileg mörk, en síðan eru setningafræðilegar reglur, **hömlur** (e. constraints), láttnar um að velja líklegustu greininguna.

Stundum getur líka dugað að greina helstu liði setningarinnar eða jafnvel einstaka hluta hennar. Í hugbúnaði til málfarsleiðréttinga er t.d. ekki síst hugað að hvers kyns samræmi, fallstjórn o.þ.h. Greining sem kölluð er **böggun** (e. chunking) dugir oft til að greina villur í þeim atriðum. Hún felst í því að greina tiltekinn setningarbút sem heild, **bagga** (e. chunk), án þess að greina tengsl þess bútar við aðra hluta setningarinnar – oftast er þarna um að ræða setningarlið en þetta getur þó einnig verið stærri heild. Til að finna – og leiðrétta – villuna í *Ég vaknaði snemma um morgunninn* er þannig nóg að greina *um morgun(n)inn* sem forsetningarlið til að vita að þarna er einu  $n$ -i ofaukið – gerð setningarinnar að öðru leyti skiptir ekki máli.

## 5. Vélræn þáttun íslenskra setninga

Grunn þáttun er að sumu leyti auðveldari í íslensku en ýmsum skyldum málum. Það stafar af því að miklar setningafræðilegar upplýsingar eru fólgnar í beygingunum. Ef við höfum greiningarforrit, **málfræðilegan markara** (e. grammatical tagger), sem greinir orðflokk og beygingarleg atriði í texta (kyn, tölu, fall, persónu, tíð, hátt, stig o.s.frv.) eru þar oft komnar nægar forsendur til að gera grein fyrir setningarliðum, hlutverki þeirra og venslum í meginatriðum.

Það er t.d. hægt að gera því skóna að nafnorð í nefnifalli sé frumlag, nema á undan því fari sögnin *vera* og þar á undan annað nafnorð – þá er seinna nafnorðið líklega sagnfylling. Nafnorð í þolfalli eða þágufalli er sennilega andlag, nema á undan því fari forsetning. Vissulega eru til aukafallsfrumlög en þau eru venjulega auðþekkt af sögnunum sem þau eiga við. Innan nafnliða er samræmi milli nafnorða, fornafna, lýsingarorða og töluorða í kyni, tölu og falli, og þau sýna oftast ytri merki um þetta samræmi og gera greiningarforriti þannig kleift að tengja þau saman sem einn lið. Svona mætti lengi halda áfram.

Í eftirfarandi setningu úr talmálsbankanum ÍSTAL (sjá Þórunni Blöndal 2005) má sjá hversu miklar setningafræðilegar upplýsingar búa í hinni málfræðilegu greiningu:

(3) Orð	Mark	Skýring
Helgi	nken-m	no (n) – kk (k) – et (e) – nf (n) – sérnafn (m)
minn	feken	fn (f) – efn (e) – kk (k) – et (e) – nf (n)
farðu	sbg2en	so (s) – bh (b) – gm – (g) – 2. pers. (2) – et (e) – nt (n)
niður	a	ao (a)
og	c	st (c)
skoðaðu	sbg2en	so (s) – bh (b) – gm – (g) – 2. pers. (2) – et (e) – nt (n)
nýja	lkeovf	lo (l) – kk (k) – et (e) – þf (o) – veik b. (v) – fst (f)
tölvuleikinn	nkeog	no (n) – kk (k) – et (e) – þf (o) – m. gr. (g)
þinn	fekeo	fn (f) – efn (e) – kk (k) – et (e) – þf (o)

Greiningarstrengurinn, **markið** (e. tag), kemur frá málfræðilegum markara sem hefur verið þjálfaður á íslenskum texta (sjá Sigrúnu Helgadóttur 2005). Hér dugir málfræðilega greiningin, mörkunin, alveg til að greina einstaka setningarliði og hlutverk þeirra; einfalt væri að skrifa forrit sem læsi markið, tengdi orð saman í liði og hengdi á liðina greiningarstrengi með heitum þeirra og hlutverkum (sjá Eirík Rögnvaldsson 2005). En vitanlega eru setningar oft talsvert flóknari en þetta og þá yrði vélræna greiningin ónákvæmari og erfiðari viðfangs.

Eins og áður segir hefur undanfarin ár verið unnið að vélrænni þáttun íslenskra setninga á vegum Friðriks Skúlasonar ehf. Þar er um að ræða djúpa þáttun sem byggist á samspili málfræðilíkansins HPSG, rökvinnskukerfisins PROLOG og beygingargreinikerfis úr *Púka* Friðriks Skúlasonar (sjá Maren Albertsdóttur og Stefán Einar Stefánsson 2004). Greiningarforritið lofar mjög góðu og getur þáttað einfaldar íslenskar setningar, en eftir er að sjá hvernig gengur að þróa það til að ráða við allan almennan texta þannig að það geti nýst í ýmiss konar notendahugbúnaði eins og lagt var upp með, s.s. málfræðileiðréttingarforritum, þýðingarforritum o.s.frv.

## 6. Lokaorð

Hér hefur verið dregið á ýmiss konar gagn sem hafa má af vélrænni þáttun setninga, t.d. í upplýsingakerfum, leiðréttingarforritum og vélrænum þýðingum. Einnig var nokkrum líkönum og aðferðum sem notuð eru við vélræna þáttun lýst lauslega, og bent á að ólík verkefni krefjast misítarlegrar greiningar. Að síðustu var vikið að vélrænni þáttun íslenskra setninga og þeim verkefnum sem unnin hafa verið á því sviði og ekki eru mörg.

Augljóslega er hér um að ræða viðfangsefni sem ekki verða gerð nein skil á fáeinum síðum. Þessum kafla er aðeins ætlað að vekja forvitni lesenda um þetta svið sem án efa á eftir að margfaldast að mikilvægi á næstu árum, og benda þeim á að hér sem víðar er ofgnótt verkefna fyrir íslenska málfræðinga framtíðarinnar.

## Lesefni

Yfirlit á íslensku um ýmsa þætti tungutækni er að finna hjá Eiríki Rögnvaldssyni (2001), en bók Jurafsky og Martin (2000) hefur að geyma ítarlega umfjöllun um öll svið tungutækni, þ. á m. setningafræðilega þáttun. Greinar eftir Maren Albertsdóttur og Stefán Einar Stefánsson (2004) og Eirík Rögnvaldsson (2005) segja svo frá tilraunum með vélræna þáttun íslenskra setninga.

## Ritaskrá

- Eiríkur Rögnvaldsson. 2001. Tungumál, tölvur og tungutækni. *Íslenskt mál* 23:7-93.
- Eiríkur Rögnvaldsson. 2005. The Corpus of Spoken Icelandic and Its Morphosyntactic Annotation. Peter Juel Henriksen (ritstj.): *Proceedings of the Seminar on Treebanks and Spoken Language, Joensuu 2005*. Copenhagen Studies in Language, Copenhagen.
- Jurafsky, Daniel, og James H. Martin. 2000. *Speech and Language Processing*. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey.
- Kaplan, Ronald, og Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. Joan Bresnan (ritstj.): *The Mental Representation of Grammatical Relations*, bls. 173-281. The MIT Press, Cambridge, Mass.
- Karlsson, Fred, Atro Voutilainen, Juha Heikkilä og Arto Anttila (ritstj.). 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
- Lønning, Jan Tore, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, Victoria Rosén og Erik Velldal. 2004. LOGON. A Norwegian MT effort. *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*. Uppsala.
- Manning, Christopher D., og Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Mass.
- Maren Albertsdóttir og Stefán Einar Stefánsson. 2004. Beygingar- og málfræðigreini-kerfi. *Sambúð tungu og tækni*, bls. 15-19. Menntamálaráðuneytið, Reykjavík.

- Pollard, Carl, og Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.
- Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. Henrik Holmboe (ritstj.): *Nordisk Sprogteknologi 2004*. Árbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004, bls. 257-265. Museum Tusulanums Forlag, København.
- Þórunn Blöndal. 2005. *Lifandi mál*. Rannsóknarstofnun Kennaraháskóla Íslands, Reykjavík.