HÁSKÓLI ÍSLANDS

Vélræn setningagreining – aðferðir, árangur og takmarkanir

Eiríkur Rögnvaldsson,

27. apríl 2007

Setningafræðileg þáttun

- Greining setningarliða
  - nafnliður, sagnliður, forsetningarliður ...
- Greining setningafræðilegra hlutverka
  - frumlag, andlag, sagnfylling ...
- Mismunandi ítarleg þáttun:
  - full þáttun (full/deep parsing)
    - heildargreining – allir möguleikar sýndir
  - hlutaþáttun (partial/shallow parsing)
    - greining í einstaka liði og setningarhlutverk

Tegundir þáttunar

- Full þáttun (full parsing; deep parsing)
  - þar sem búið er til fullkomið **þáttunartré** (parse tree) fyrir sérhverja setningu
  - oft margir möguleikar
- Hlutaþáttun (partial parsing; shallow parsing)
  - þar sem setningar eru greindar í setningarhluta
  - án þess að krefjast þess að sérhver hluti passi inn í víðtæka þáttun (e. global parse)

Þáttun kemur víða að gagni

- Málfræðileiðrétting í ritvinnslukerfum
  - ef ekki er hægt að þátta setningu bendir það til að í henni séu villur – eða hún sé a.m.k.strembin
- Merkingargreining
  - nýtist í vélrænum þýðingum
  - sjálfvirkri svörun
  - útdrætti upplýsinga
- Talkennsl

Þáttunarferlið

- Full þáttun felst í leit
  - farið er gegnum öll hugsanleg þáttunartré
  - til að finna það sem á við setninguna
- Leitin stýrist af tvennu
  - annars vegar ílaginu, þ.e. orðunum í setningunni
  - hins vegar málfræðinni
- Fundin öll tré með rótina *S*
  - og orðin í setningunni sem lokatákn

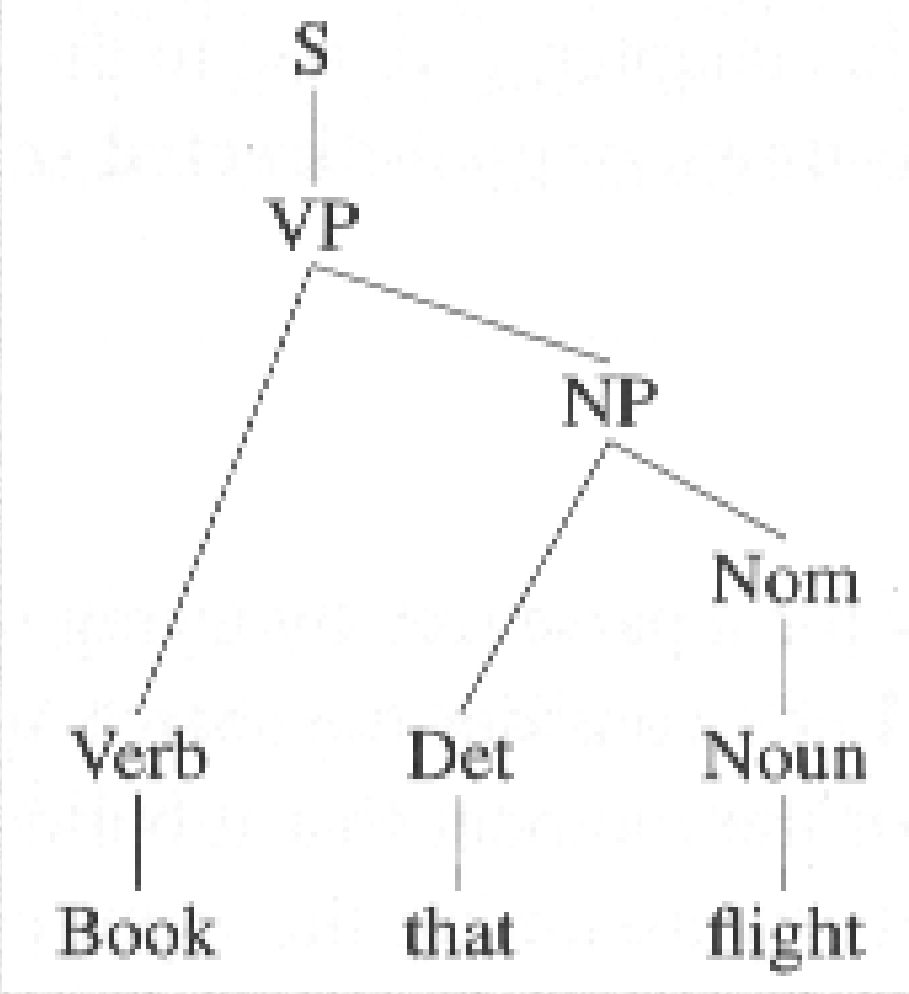
Brot úr enskri málfræði

- Fáeinar einfaldar liðgerðarreglur úr ensku

$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	
$Nominal \rightarrow Noun\ Nominal$	$Prep \rightarrow from \mid to \mid on$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	$Nominal \rightarrow Nominal\ PP$

Þáttun einfaldrar setningar

- Ein greining möguleg
  - miðað við gefnar reglur
    - S → VP
    - VP → Verb NP
    - NP → Det Nominal
    - Nominal → Noun
    - Verb → *book*
    - Det → *that*
    - Noun → *flight*



Ofansækni og neðansækni

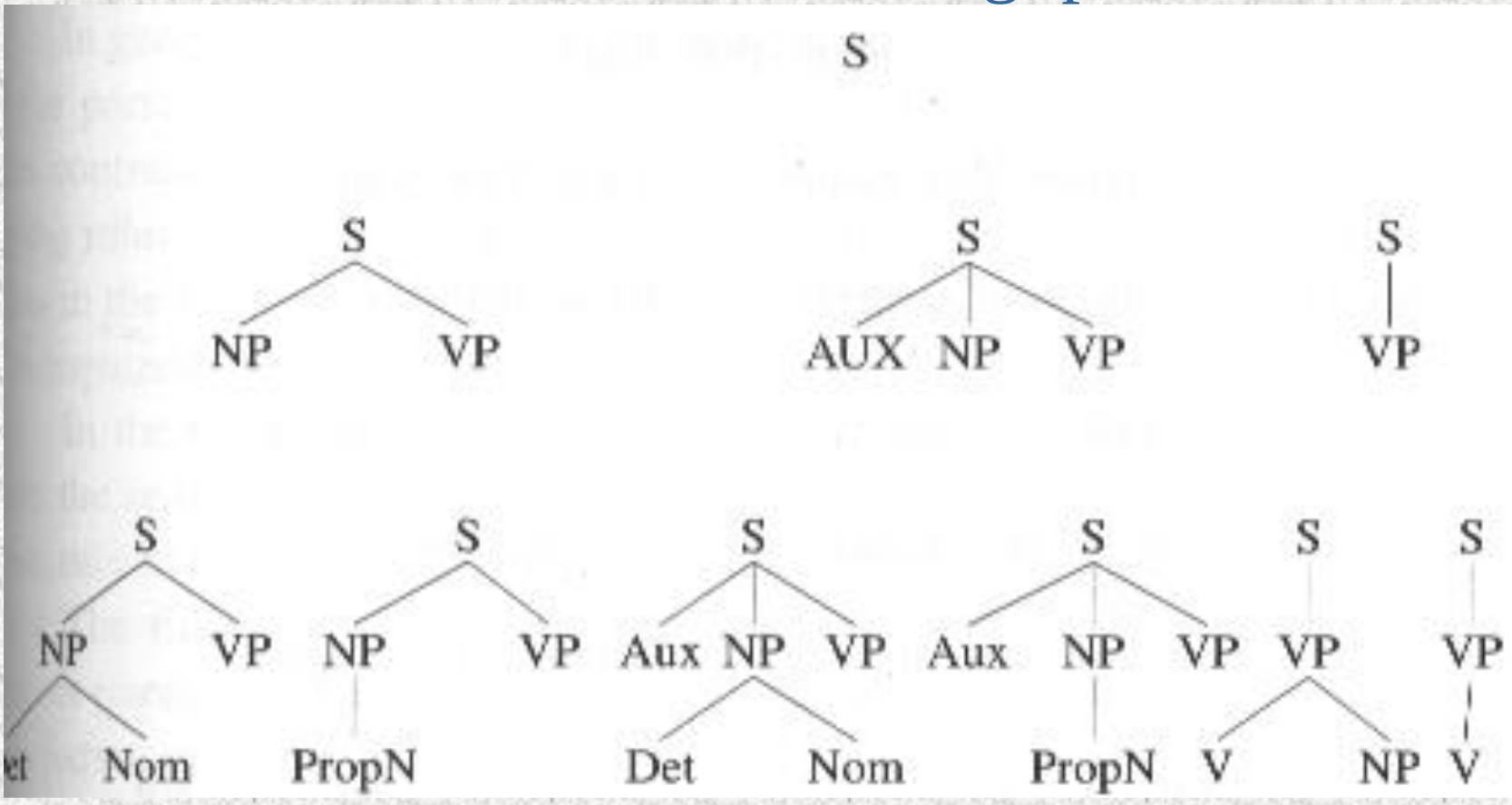
- Tvenns konar leitaraðferðir
  - tvær mismunandi áttir
- Ofansækin leit (top-down, goal-directed)
  - byrjar á *S* efst í trénu
  - og leitar niður á við, að orðunum
- Neðansækin leit (bottom-up, data-directed)
  - byrjar á orðunum
  - og leitar upp á við, allt að *S*

Ofansækin þáttun

- Byrjað á *S*
  - búnir til allir trjátoppar sem byrja á *S*
  - út frá öllum reglum með *S* vinstra megin við ör
- Síðan er haldið áfram í næsta lagi
  - tekin öll tákn næst neðan við *S*
  - fundnar reglur þar sem þau eru vinstra megin
- Þannig er haldið áfram þar til kemur að orðum
  - þá er trjám sem ekki passa við orðin hent

Dæmi um ofansækna þáttun

- Aðeins næstsíðasta tréð í 3. lagi passar

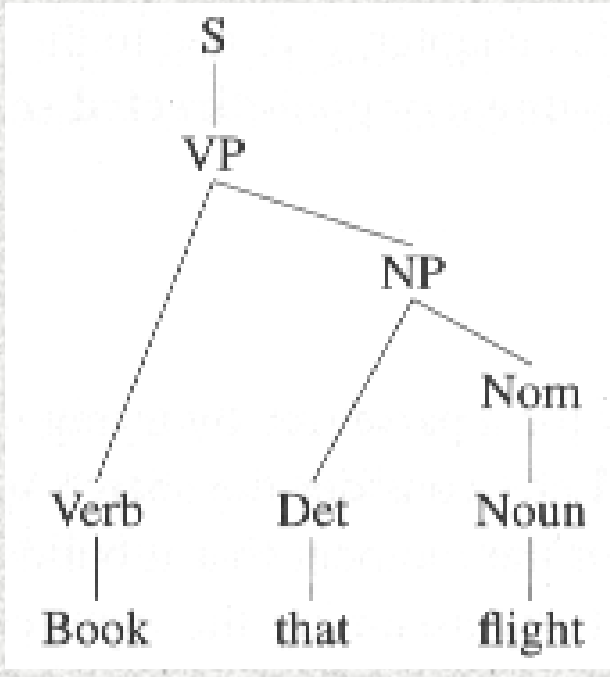


Neðansækin þáttun

- Byrjað á orðunum
  - þeim er flett upp í orðasafni
  - og skrifaðir út hugsanlegir orðflokkar
- Síðan er reynt að tengja orðin saman
  - leitað að strengjum sem passa við það sem er hægra megin við örina í einhverri reglu
- Þannig er haldið áfram þar til kemur að *S*
  - trjám sem ekki leiða þangað er hent

Dæmi

- Neðansækin þáttun
  – endanlegt tré ↓



Kostir og gallar aðferðanna

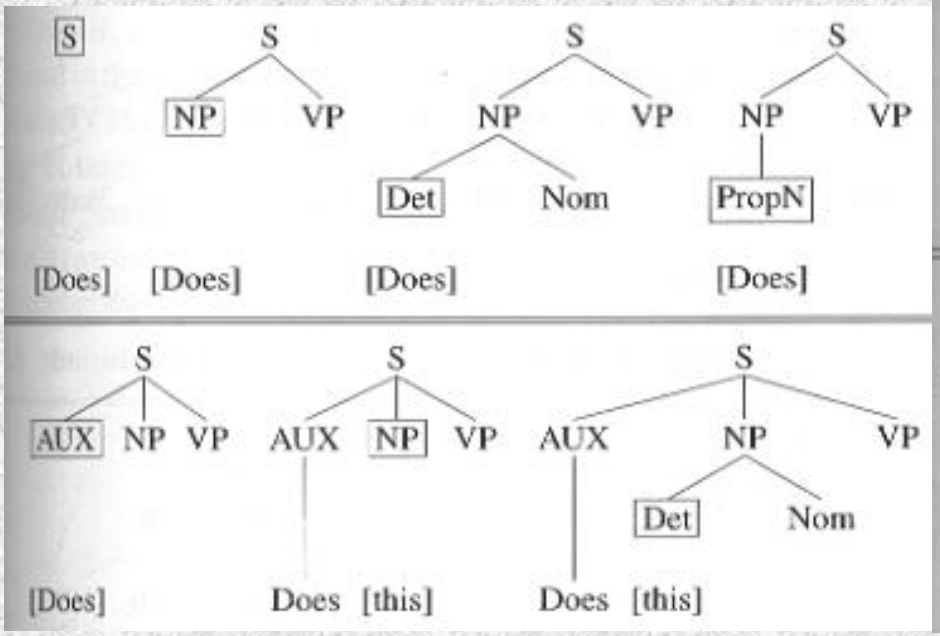
- Ofansækin þáttun gengur út frá *S*
  - eyðir aldrei tíma í tré sem ekki geta endað í *S*
- Hún horfir ekki á orðin í setningunni
  - býr því til fjölda trjáa sem ekki falla að gögnunum
- Neðansækin þáttun gengur út frá orðunum
  - eyðir ekki tímanum í tré sem ekki falla að gögnum
- Hún horfir ekki á rótina *S*
  - býr því til trjábúta sem aldrei geta orðið að heilu tré

Breidd og dýpt

- Hvernig á að fara í gegnum möguleikana?
  - nota breiddarleit (breadth-first)
  - eða dýptarleit (depth-first)?
- Hér er valin dýptarleit
  - einn möguleiki á hverju sviði valinn
  - ef hann bregst er farið á næsta svið á undan
  - og annar möguleiki þar valinn til skoðunar

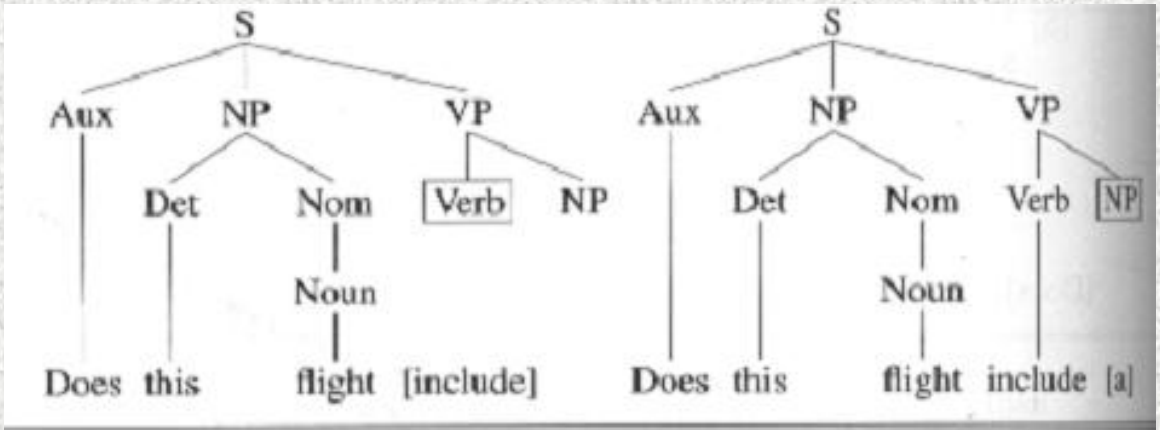
Þáttun hefst

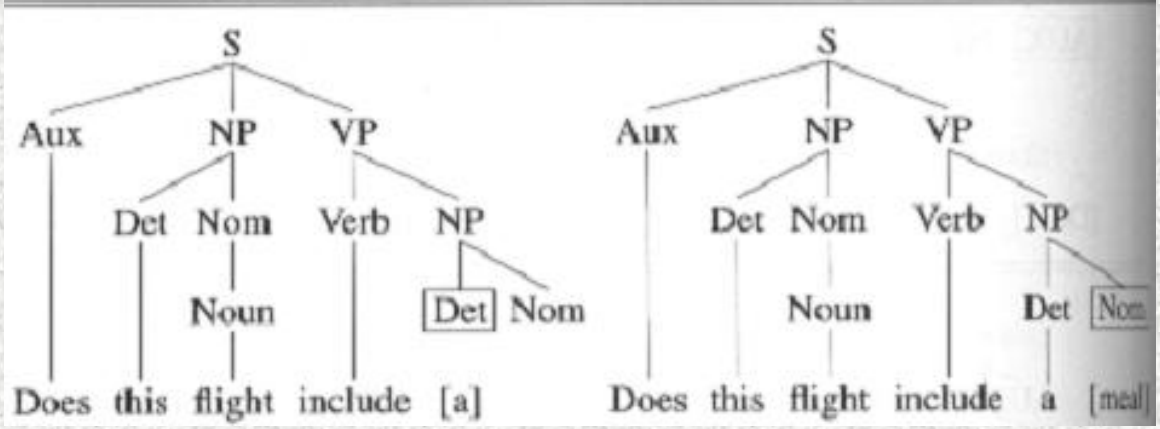
- Hér er sýnt hvernig farið er gegnum setninguna *Does this flight include a meal?*

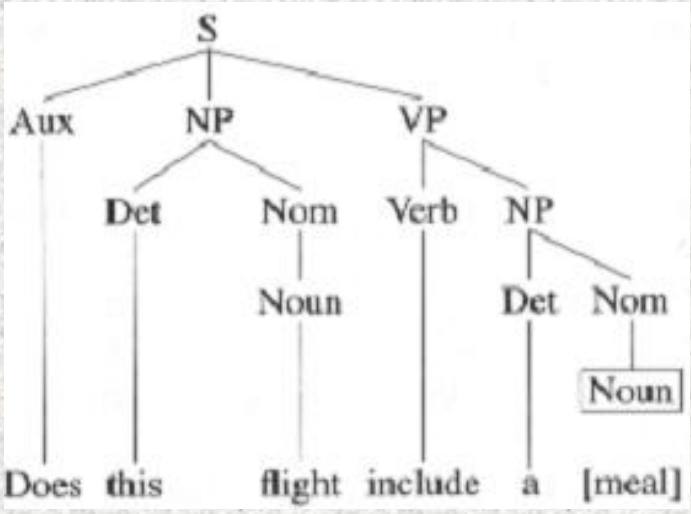


Þáttun tekst

- Hér er haldið áfram uns þáttun tekst







Neðansækin síun

- Þegar þáttunin kemur niður að orði
  - er skoðað hvort það geti fallið inn í tréð
  - sé svo er næsta orð tekið til skoðunar
- Það orð verður þá að geta verið fyrsta orð
  - í reglu sem á við næstu eind á sviðinu fyrir ofan
  - annars er greining orðsins á undan endurskoðuð
  - þannig er komið í veg fyrir að prófaðir séu möguleikar sem augljóslega ganga ekki upp

Hlutaþáttun

- Í mörgum tilvikum er nægjanlegt að greina setningar í setningarhluta eða setningarliði
  - án þess að krefjast þess að liðirnir passi inn í víðtækt þáttunartré
- Þetta getur átt við á ýmsum sviðum
  - **upplýsingaútdrætti** (e. information extraction)
  - eða **textaútdrætti** (e. text summarization)
  - þar sem greining setningarliða er mikilvægari en full þáttun

Mismunandi þáttun setningar

- Full þáttun – mismunandi greiningar:
  - $[_{S}\ [_{NL}$ Margir$]\ [_{SL}$ hittu $[_{NL}$ Maríu $[_{FL}$ á $[_{NL}$ skrifstofunni$]]]]]$
    - eða:
  - $[_{S}\ [_{NL}$ Margir$]\ [_{SL}$ hittu $[_{NL}$ Maríu$]]\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]]$
- Hlutaþáttun – ein greining:
  - $\{_{FRL}\ [_{NL}$ Margir$]\}\ [_{SL}$ hittu$]\ \{_{ANDL}\ [_{NL}$ Maríu$]\}\ [_{FL}$ á $[_{NL}$ skrifstofunni$]]$
- Setningarliðirnir ekki felldir saman í eitt tré

Kostir hlutaþáttunar

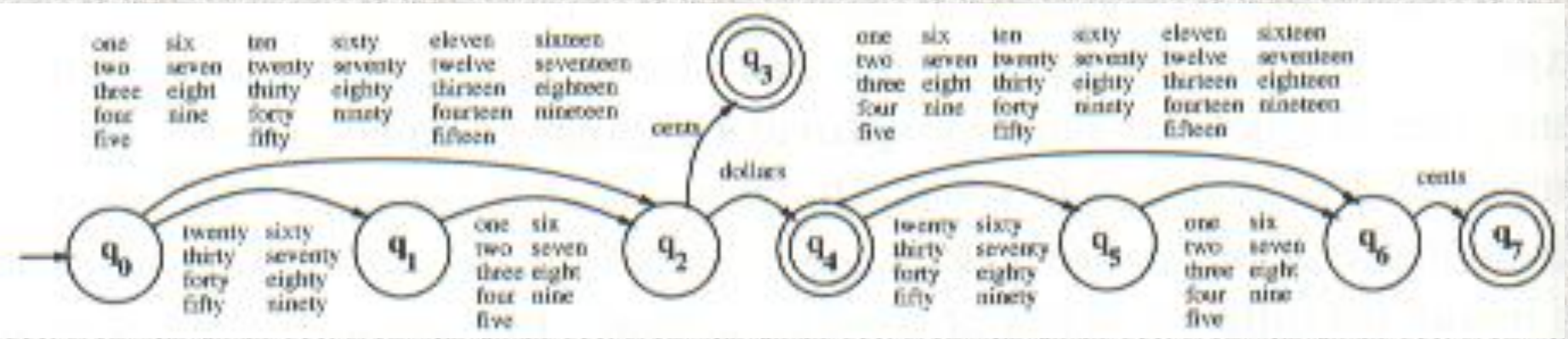
- Full þáttun
  - nákvæmari og sýnir alla möguleika, en:
  - frek á tíma og reiknigetu
  - viðkvæm fyrir villum í inntaki
- Hlutaþáttun
  - sýnir ekki formgerðina eins nákvæmlega, en:
  - skilar greiningu þrátt fyrir villur í inntaki
  - hentar því vel t.d. fyrir texta á netinu

Upphaflegur texti

- pabbi nken hennar fpvee var sfg3eþ vinur nken minn feken og c við fp1fn teflum sfg1fþ oft aa saman aa á ao kvöldin nhfog og c hann fpken studdi sfg3eþ mig fp1eo og c ég fp1en hélt sfg1eþ meira lheovm að cn segja sng áfram aa að cn koma sng til ae hans fpkee eftir ao að c allt fohen var sfg3eþ orðið sþghen skrýtið lhensf og c hún fpven hætt sþgven að cn hringja sng í ao mig fp1eo . .

Stöðuvélaaðferð

- Hér er notuð stöðuvélaaðferð
  - finite state method



- Aðferðin hefur ýmsa kosti
  - en ræður t.d. ekki við endurkvæmni

Orðasambönd

- pabbi nken hennar fpvee var sfg3eþ vinur nken minn feken og c við fp1fn teflum sfg1fþ oft aa saman aa á ao kvöldin nhfog og c hann fpken studdi sfg3eþ mig fp1eo og c ég fp1en hélt sfg1eþ *[MWE_AdvP meira lheovm að cn segja sng MWE_AdvP]* áfram aa að cn koma sng til ae hans fpkee *[MWE_SCP eftir ao að c MWE_SCP]* allt fohen var sfg3eþ orðið sþghen skrýtið lhensf og c hún fpven hætt sþgven að cn hringja sng í ao mig fp1eo . .

Atviksliðir og tengingar

- pabbi nken hennar fpvee var sfg3eþ vinur nken minn feken *[CP og c CP]* við fp1fn teflum sfg1fþ *[AdvP oft aa AdvP] [AdvP saman aa AdvP]* á ao kvöldin nhfog *[CP og c CP]* hann fpken studdi sfg3eþ mig fp1eo *[CP og c CP]* ég fp1en hélt sfg1eþ [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] *[AdvP áfram aa AdvP]* að cn koma sng til ae hans fpkee [MWE_SCP eftir ao að c MWE_SCP] allt fohen var sfg3eþ orðið sþghen skrýtið lhensf *[CP og c CP]* hún fpven hætt sþgven að cn hringja sng í ao mig fp1eo . .

Lýsingarorðsliðir

- pabbi nken hennar fpvee var sfg3eþ vinur nken minn feken [CP og c CP] við fp1fn teflum sfg1fþ [AdvP oft aa AdvP] [AdvP saman aa AdvP] á ao kvöldin nhfog [CP og c CP] hann fpken studdi sfg3eþ mig fp1eo [CP og c CP] ég fp1en hélt sfg1eþ [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] að cn koma sng til ae hans fpkee [MWE_SCP eftir ao að c MWE_SCP] allt fohen var sfg3eþ orðið sþghen *[AP skrýtið lhensf AP]* [CP og c CP] hún fpven hætt sþgven að cn hringja sng í ao mig fp1eo . .

Nafnliðir

- *[NP pabbi nken NP] [NP hennar fpvee NP]* var sfg3eþ *[NP vinur nken minn feken NP]* [CP og c CP] *[NP við fp1fn NP]* tefldum sfg1fþ [AdvP oft aa AdvP] [AdvP saman aa AdvP] á ao *[NP kvöldin nhfog NP]* [CP og c CP] *[NP hann fpken NP]* studdi sfg3eþ *[NP mig fp1eo NP]* [CP og c CP] *[NP ég fp1en NP]* hélt sfg1eþ [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] að cn koma sng til ae *[NP hans fpkee NP]* [MWE_SCP eftir ao að c MWE_SCP] *[NP allt fohen NP]* var sfg3eþ orðið sþghen [AP skrýtið lhensf] [CP og c CP] *[NP hún fpven NP]* hætt sþgven að cn hringja sng í ao *[NP mig fp1eo NP]* . .

Sagnir

- [NP pabbi nken NP] [NP hennar fpvee NP] *[VPb var sfg3eþ VPb]* [NP vinur nken minn feken NP] [CP og c CP] [NP við fp1fn NP] *[VP tefldum sfg1fþ VP]* [AdvP oft aa AdvP] [AdvP saman aa AdvP] á ao [NP kvöldin nhfog NP] [CP og c CP] [NP hann fpken NP] *[VP studdi sfg3eþ VP]* [NP mig fp1eo NP] [CP og c CP] [NP ég fp1en NP] *[VP hélt sfg1eþ VP]* [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] *[VPi að cn koma sng VPi]* til ae [NP hans fpkee NP] [MWE_SCP eftir ao að c MWE_SCP] [NP allt fohen NP] *[VPb var sfg3eþ VPb] [VPp orðið sþghen VPp]* [AP skrýtið lhensf] [CP og c CP] [NP hún fpven NP] *[VPp hætt sþgven VPp] [VPi að cn hringja sng VPi]* í ao [NP mig fp1eo NP] . .

Forsetningarliðir

- [NPn pabbi nken] [NPg hennar fpvee] [VPb var sfg3eþ VPb] [NPn vinur nken minn feken] [CP og c CP] [NPn við fp1fn] [VP tefldum sfg1fþ VP] [AdvP oft aa AdvP] [AdvP saman aa AdvP] *[PP á ao [NPa kvöldin nhfog] PP]* [CP og c CP] [NPn hann fpken] [VP studdi sfg3eþ VP] [NPa mig fp1eo] [CP og c CP] [NPn ég fp1en] [VP hélt sfg1eþ VP] [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] [VPi að cn koma sng VPi] *[PP til ae [NPg hans fpkee] PP]* [MWE_SCP eftir ao að c MWE_SCP] [NPn allt fohen] [VPb var sfg3eþ VPb] [VPp orðið sþghen VPp] [APn skrýtið lhensf] [CP og c CP] [NPn hún fpven] [VPp hætt sþgven VPp] [VPi að cn hringja sng VPi] *[PP í ao [NPa mig fp1eo] PP]* . .

Eignarfallseinkunnir

- [NPn pabbi nken] *{*QUAL [NPg hennar fpvee] *QUAL}* [VPb var sfg3eþ VPb] [NPn vinur nken minn feken] [CP og c CP] [NPn við fp1fn] [VP teflum sfg1fþ VP] [AdvP oft aa AdvP] [AdvP saman aa AdvP] [PP á ao [NPa kvöldin nhfog] PP] [CP og c CP] [NPn hann fpken] [VP studdi sfg3eþ VP] [NPa mig fp1eo] [CP og c CP] [NPn ég fp1en] [VP hélt sfg1eþ VP] [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] [VPi að cn koma sng VPi] [PP til ae *{*QUAL [NPg hans fpkee] *QUAL}* PP] [MWE_SCP eftir ao að c MWE_SCP] [NPn allt fohen] [VPb var sfg3eþ VPb] [VPp orðið sþghen VPp] [APn skrýtið lhensf] [CP og c CP] [NPn hún fpven] [VPp hætt sþgven VPp] [VPi að cn hringja sng VPi] [PP í ao [NPa mig fp1eo] PP] . .

Frumlög

- *{*SUBJ> [NPn pabbi nken] {*QUAL [NPg hennar fpvee] *QUAL} *SUBJ>}* [VPb var sfg3eþ VPb] [NPn vinur nken minn feken] [CP og c CP] *{*SUBJ> [NPn við fp1fn] *SUBJ>}* [VP tefldum sfg1fþ VP] [AdvP oft aa AdvP] [AdvP saman aa AdvP] [PP á ao [NPa kvöldin nhfog] PP] [CP og c CP] *{*SUBJ> [NPn hann fpken] *SUBJ>}* [VP studdi sfg3eþ VP] [NPa mig fp1eo] [CP og c CP] *{*SUBJ> [NPn ég fp1en] *SUBJ>}* [VP hélt sfg1eþ VP] [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] [VPi að cn koma sng VPi] [PP til ae {*QUAL [NPg hans fpkee] *QUAL} PP] [MWE_SCP eftir ao að c MWE_SCP] *{*SUBJ> [NPn allt fohen] *SUBJ>}* [VPb var sfg3eþ VPb] [VPp orðið sþghen VPp] [APn skrýtið lhensf] [CP og c CP] *{*SUBJ [NPn hún fpven] *SUBJ}* [VPp hætt sþgven VPp] [VPi að cn hringja sng VPi] [PP í ao [NPa mig fp1eo] PP] . .

Andlög og sagnfyllingar

- {*SUBJ> [NPn pabbi nken] {*QUAL [NPg hennar fpvee] *QUAL} *SUBJ>} [VPb var sfg3eþ VPb] *{*COMP< [NPn vinur nken minn feken] *COMP<}* [CP og c CP] {*SUBJ> [NPn við fp1fn] *SUBJ>} [VP tefldum sfg1fþ VP] [AdvP oft aa AdvP] [AdvP saman aa AdvP] [PP á ao [NPa kvöldin nhfog] PP] [CP og c CP] {*SUBJ> [NPn hann fpken] *SUBJ>} [VP studdi sfg3eþ VP] *{*OBJ< [NPa mig fp1eo] *OBJ<}* [CP og c CP] {*SUBJ> [NPn ég fp1en] *SUBJ>} [VP hélt sfg1eþ VP] [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP] [AdvP áfram aa AdvP] [VPi að cn koma sng VPi] [PP til ae {*QUAL [NPg hans fpkee] *QUAL} PP] [MWE_CP eftir ao að c MWE_CP] {*SUBJ> [NPn allt fohen] *SUBJ>} [VPb var sfg3eþ VPb] *{*COMP< [VPp orðið sþghen VPp] *COMP<}* *{*COMP [APn skrýtið lhensf] *COMP}* [CP og c CP] {*SUBJ [NPn hún fpven] *SUBJ} *{*COMP [VPp hætt sþgven VPp] *COMP}* [VPi að cn hringja sng VPi] [PP í ao [NPa mig fp1eo] PP] . .

Lokaútkoma - 1

- {*SUBJ> [NP pabbi nken NP] {*QUAL [NP hennar fpvee NP] *QUAL} *SUBJ>}
- [VPb var sfg3eþ VPb]
- {*COMP< [NPn vinur nken minn feken] *COMP<}
- [CP og c CP]
- {*SUBJ> [NPn við fp1fn] *SUBJ>}
- [VP tefldum sfg1fþ VP]
- [AdvP oft aa saman aa AdvP]
- [PP á ao [NP kvöldin nhfog NP] PP]
- [CP og c CP]
- {*SUBJ> [NP hann fpken NP] *SUBJ>}
- [VP studdi sfg3eþ VP]
- {*OBJ< [NP mig fp1eo NP] *OBJ<}
- [CP og c CP]
- {*SUBJ> [NP ég fp1en NP] *SUBJ>}
- [VP hélt sfg1eþ VP]

Lokaútkoma - 2

- [MWE_AdvP meira lheovm að cn segja sng MWE_AdvP]
- [AdvP áfram aa AdvP]
- [VPi að cn koma sng VPi]
- [PP til ae [NP hans fpkee NP] PP]
- [MWE_SCP eftir aþ að c MWE_SCP]
- {*SUBJ> [NP allt fohen NP] *SUBJ>}
- [VPb var sfg3eþ VPb]
- {*COMP< [VPp orðið sþghen VPp] *COMP<}
- {*COMP< [AP skrýtið lhensf AP] *COMP<}
- [CP og c CP]
- {*SUBJ [NP hún fpven NP] *SUBJ}
- {*COMP [VPp hætt sþgven VPp] *COMP}
- [VPi að cn hringja sng VPi]
- [PP í ao [NP mig fp1eo NP] PP]
- . .

Mat á frammistöðu þáttarans

- Búið var til prófunarsafn
  - 509 setningar úr grunni *Íslenskrar orðtíðnibókar*
  - valdar tilviljanakennt
- Þetta safn var greint í höndunum
  - í samræmi við þáttunarskemað
- Sú greining myndar *gold standard*
  - sem greining þáttarans er borin saman við

Setningarliðir

- Nákvæmni í greiningu setningarliða
  - bæði miðað við rétt mörk úr *Íslenskri orðtíðnibók* og mörk úr *IceTagger*

Phrase type	F-measure correct tags	F-measure *IceTagger*	Freq. in test data
AdvP	91.8%	85.1%	8.2%
AP	95.1%	86.3%	8.1%
APs	87.0%	68.6%	0.5%
NP	96.8%	93.0%	37.6%
NPs	80.4%	74.3%	1.5%
PP	96.7%	91.3%	13.0%
VPx	99.2%	93.8%	19.3%
CP	100.0%	99.6%	5.7%
SCP	99.6%	97.6%	3.4%
InjP	100.0%	96.3%	0.2%
MWE	96.9%	92.6%	2.5%
All	96.7%	91.9%	100.0%

Setningafræðileg hlutverk

- Nákvæmni í greiningu setninga-fræðilegra hlutverka
  - bæði miðað við rétt mörk úr *Íslenskri orðtíðnibók* og mörk úr *IceTagger*

Function type	F-measure correct tags	F-measure *IceTagger*	Freq. in test data
SUBJ	68.2%	47.6%	4.7%
SUBJ>	92.7%	89.4%	30.3%
SUBJ<	83.7%	75.1%	12.3%
OBJ	0.0%	0.0%	0.2%
OBJ>	43.5%	20.0%	0.8%
OBJ<	90.2%	78.2%	19.7%
OBJAP>	71.4%	57.2%	0.2%
OBJAP<	75.0%	46.2%	0.4%
OBJNOM<	30.8%	16.7%	0.6%
IOBJ<	73.3%	51.9%	0.9%
COMP	56.9%	40.0%	2.8%
COMP>	91.3%	91.3%	1.3%
COMP<	75.1%	70.0%	12.7%
QUAL	87.7%	77.9%	10.4%
TIMEX	74.7%	55.9%	2.7%
All	84.3%	75.3%	100.0%

Röng greining liða, 1

- Ranglega greindur atviksliður:
  - [PP um [NP það NP] PP] [VP vissi VP] [NP stelpan NP] **[AdvP ekki þá AdvP]**
    - hér er *ekki* setningaratviksorð en stendur ekki með tíðaratviksorðinu *þá*
- Ranglega greindur lýsingarorðsliður:
  - [CP og CP] [VP tóku VP] [NP **[AP [AdvP fram AdvP] eigin AP]** dósir NP]
    - hér er *fram* sagnarögn en stendur ekki með *eigin*

Röng greining liða, 2

- Ranglega greindur samsettur nafnliður:
  - [AP sterkur AP] [VPb var VPb] **[NPs [NP hann NP] [CP og CP] [NP íþróttamaður NP] NPs]** [AP ágætur AP]
    - hér standa *hann* og *íþróttamaður* í sama falli og aðaltenging á milli, og eru því greindir sem samsettur nafnliður
    - á hinn bóginn er lo. *ágætur* ekki greint sem hluti nafnliðar af því að það stendur á eftir no.

Röng greining hlutverka

- Ófullkomin greining frumlags
  - [VPb er VPb] [AdvP ekki AdvP] [VPi að koma VPi] {*SUBJ [NP matur NP] *SUBJ}?
    - hér truflar liður milli sagnar og frumlags greininguna
- Röng liðgreining > röng hlutverksgreining
  - {*OBJ< [NP [AP [AdvP fram AdvP] eigin AP] dósir NP] *OBJ<}
    - *fram* greint sem hluti andlags af því að það var greint sem hluti lýsingarorðsliðar

Niðurstöður

- Niðurstöðurnar lofa góðu
  - 96,7% nákvæmni í greiningu setningarliða
  - 84,3% nákvæmni í greiningu hlutverka
- Árangurinn mætti bæta með því að
  - nýta beygingarlegar upplýsingar meira
  - byggja meira á ýmiss konar orðalistum
  - endurbæta stöðuferjöldin og fjölga þeim