

# Vélræn þáttun forníslensku: Samanburður aðferða

**Lykilorð:** þáttun, trjábanki, forníslenska, setningafræði

## Efnisyfirlit

<b>1</b>	<b>Leiðbeiningar</b>	<b>1</b>
1.1	Grammar búið til . . . . .	1
1.2	Texti þáttaður . . . . .	2
1.3	Um tölfræðiþáttarann . . . . .	3
	<b>Heimildaskrá</b>	<b>3</b>

## 1 Leiðbeiningar

### 1.1 Grammar búið til

Til þess að búa til grammar þarf að hafa nýjustu útgáfuna af BerkeleyParser.jar (þegar þetta er skrifað er það BerkeleyParser1-7.jar). Því næst er farið í skipanaglugga (command-glugga). Hér er training.psd skjal með þeim trjám sem á að æfa á. ice.gr er svo grammar-fællinn sem verður búinn til.

- (1) `java -cp BerkeleyParser-1.7.jar edu.berkeley.nlp.PCFGGLA.GrammarTrainer -path simpleicepahc7selectedtextswithouttextitexti.psd -out ice.gr -treebank SINGLE-FILE`

Þegar textarnir voru valdir var öðrum hverjum texta hent út. Textarnir sem ice.gr var æft á eru: 1150.homiliubok, 1210.thorlakur, 1250.thetubrot, 1270.gragas, 1300.alexander, 1325.arni, 1350.finnbogi, 1400.gunnar, 1400.viglundur, 1450.ectorssaga, 1450.vilhjalmur, 1480.jarlmann, 1525.georgius, 1540.ntjohn, 1611.okur, 1630.gerhard, 1659.pislarsaga, 1675.armann, 1675.modars, 1720.vidalin, 1745.klim, 1791.jonsteingrims, 1835.jonasedli, 1859.hugvekjur, 1882.torfhildur, 1888.grimur, 1902.fossar, 1908.ofurefli, 1985.margsaga, 2008.mamma

Textarnir sem síðan eru notaðir til að athuga hve góður þáttarinn ice.gr er eru: 1150.firstgrammar, 1210.jartein, 1250.sturlunga, 1260.jomsvikingar, 1275.morkin, 1310.grettir, 1350.bandamennM, 1350.marta, 1400.gunnar2, 1450.bandamenn, 1450.judit, 1475.aevintyri, 1525.erasmus, 1540.ntacts, 1593.eintal, 1628.olafuregils, 1650.illugi, 1661.indiafari, 1675.magnus, 1680.skalholt, 1725.biskupasogur, 1790.fimmbraedra, 1830.hellismenn, 1850.piltur, 1861.orrusta, 1883.voggur, 1888.vordraumur, 1907.leysing, 1920.arin, 1985.sagan, 2008.ofsi

## 1.2 Texti þáttaður

Byrjað er á því að breyta trjábankanum (ath. hér er átt við útgefinn trjábanka, IcePaHC 0.9 en ekki nýjustu gerð hans á github), sem inniheldur þáttaða texta, í óþáttaðar textaskrár, í þeirri röð sem hér er sýnt:

- (2) a. `./psd-to-pos fæll.psd [gefur .pos fæl]`  
b. `python3 rm-lemmata.py fæll.pos > fæll1.psd`  
c. `./psd-to-pos-post-rmlemmata fæll1.psd [gefur .pos fæl]`  
d. `python3 rm-lemmata-posteverything.py fæll1.pos > fæll2.psd`

Því næst eru textaskrárnar sem búnar eru til með þessum hætti þáttaðar með tölfræðipáttaranum. Það er gert með eftirfarandi skipun (hér er `fæll2.psd` inntaksskráin og `fæll2parsed` úttakið, þ.e. skrá með textanum þáttaðum):

- (3) `java -jar BerkeleyParser-1.7.jar -gr icy.gr -inputFile fæll2.psd -outputFile fæll2parsed.tok`

Að þátta um 20.000 orð tekur u.þ.b. 15 mínútur á tölvu með venjulegum örgjörva.

Næsta skref er að láta IceTagger marka textann. Hér þarf að beita „svindli“ vegna þess að markarinn lítur á svigamerki sem mörk sem þarf að marka rétt eins og heiti á liðum og orðflokkum (svo sem CP eða NOUN). Svindlið felst í því að breyta skilgreiningum í `encodemarkup.py` og `decodemarkup.py`.

- (4) `./txt2ipsd.sh python3 fæll2parsed [tekur .tok fæl og skilar honum mörkuðum; vinnur með encodemarkup.py og decodemarkup.py]`

Þegar þessu er lokið er unnið með skrána `tag-word.py` sem breytir röð orða og marka, þannig að mark fari næst á undan viðkomandi orði en ekki næst á eftir (eins og IceTagger gerir).

- (5) `python3 tag-word.py fæll2 > fæll2.psd`

Því næst er unnið með skrána X sem breytir mörkunum í IcePaHC-mörk, þ.e. mörk sem eru sams konar og þau sem voru notuð í IcePaHC.

- (6)

Að endingu er svo unnið með skrána Y sem eyðir bilum þannig að hægt sé að keyra CorpusSearch á skrárnar.

- (7) `((IP (NP-PRD (nken-m Grímur)) (sfg3ep hét) (NP-SBJ (lkensf einn) (nken bóndi) (. .) (NP (lheesf mikils) (nkee háttar) (CONJP (c og) (ADJP (aa vel) (lkenof fjáreigandi)))))) (. .)))`

## 1.3 Um tölfræðipáttarann

Möguleg atriði:

- Orðaröð — VO, OV
- Nafnliðurinn — innbyrðis röð nafnorða, lýsingarorða, magnorða
- Röð óbeins andlags og beins andlags
- Forsetningarstrand — niðurstaðan verður slæm fyrir tölfræðipáttarann
- Kjarnafærsla í aukasetningum — niðurstaðan gæti verið slæm
- Ákveðnihamlan — líklega ekki mjög heppilegt
- Ópersónulega háttarsagnagerðin
- Frásagnarumröðun — ætti að vera fínt

## Heimildaskrá