# Sögulegur íslenskur trjábanki og nýting hans

Eiríkur Rögnvaldsson
Anton Karl Ingason
Einar Freyr Sigurðsson
Joel Wallenberg

# Útdráttur

- Í haust lauk gerð Sögulegs íslensks trjábanka (Icelandic Parsed Historical Corpus, IcePaHC), safns setningafræðilegra greindra texta frá 12. til 21. aldar – alls ein milljón orða úr rúmum 60 textum. Frumgreining textanna var vélræn en meginvinnan við bankann fólst í handvirkri greiningu textanna sem var mikið verk.
- Trjábankinn er gerður í tvennum tilgangi. Annars vegar til nota í máltækni, en nákvæmar upplýsingar um setningagerð eru mikilvæg forsenda fyrir gerð ýmiss konar máltæknibúnaðar, svo sem leiðréttingarforrita, vélrænna þýðinga o.fl. Hins vegar er bankinn ætlaður til málrannsókna, einkum á setningagerð og setningafræðilegum breytingum, og hefur þegar sannað gildi sitt í ýmsum rannsóknum af því tagi.

# Íslenski trjábankinn og sérstaða hans

- Tvenns konar nýting
  - ætlaður bæði fyrir máltækni og málfræðirannsóknir
- Spönnun í tíma
  - rúm 800 ár – varla raunhæft í öðrum tungumálum
- Stærð
  - ein milljón orða – aðeins tveir trjábankar stærri
- Aðgengi
  - algerlega opinn og ókeypis, notkun án takmarkana

# Styrktaraðilar

- Rannsóknasjóður, öndvegisstyrkur:
  - Hagkvæm máltækni utan ensku - íslenska tilraunin
- U.S. National Science Foundation (NSF):
  - Evolution of Language Systems: a comparative study of grammatical change in Icelandic and English
- Rannsóknasjóður Háskóla Íslands:
  - Sögulegur íslenskur trjábanki
- ICT Policy Support Programme:
  - META-NORD

# Aðstandendur

- Verkefnisstjórar:
  - Eiríkur Rögnvaldsson
  - Joel C. Wallenberg
- Hönnuðir og þáttarar:
  - Anton Karl Ingason
  - Einar Freyr Sigurðsson
  - Joel C. Wallenberg
    - Brynhildur Stefánsdóttir
    - Hulda Óladóttir
- IceNLP:
  - Hrafn Loftsson
- Erlendir samstarfsaðilar:
  - Tony Kroch (UPenn)
  - Beatrice Santorini (UPenn)
- Aðrir:
  - Ýmis forlög, útgefendur og fræðimenn útveguðu texta
  - Nokkrir rithöfundar veittu leyfi til notkunar texta
  - Nokkrir stúdentar unnu við innslátt

# Upphafleg áætlun

- Fimm textaflokkar
  - Frásagnartextar (nar)
  - Trúarlegir textar (rel)
  - Ævisögur og ferðabækur (bio)
  - Fræðitextar (sci)
  - Lagatextar, dómar, skjöl (law)
- 20 þúsund orð af hverri tegund frá hverri öld
  - alls 100 þúsund orð á öld
  - ein milljón orða í heild

# Textaöflun

- Þetta reyndist óframkvæmanlegt að sinni
  - útilokað að fá texta í öllum flokkum frá öllum öldum
- Ákveðið að leggja áherslu á frásagnartexta
  - þeir eru rúmlega 2/3 af heildinni
    - frá miðri 19. öld er um skáldsögur að ræða
  - trúartextar eru tæpur fjórðungur
    - frá öllum öldum nema 15. og 21.
  - ævisögur og ferðabækur um 75 þúsund orð
    - annað óverulegt

# Textategundir og orðafjöldi

| | nar | rel | bio | sci | law | Total |
|---|---|---|---|---|---|---|
| 12th | 0 | 40871 | 0 | 4439 | 0 | **45310** |
| 13th | 93463 | 21196 | 0 | 0 | 6183 | **120842** |
| 14th | 77370 | 21315 | 0 | 0 | 0 | **98685** |
| 15th | 111560 | 0 | 0 | 0 | 0 | **111560** |
| 16th | 35733 | 60464 | 0 | 0 | 0 | **96197** |
| 17th | 46281 | 28134 | 52997 | 0 | 0 | **127412** |
| 18th | 63322 | 22963 | 22099 | 0 | 0 | **108384** |
| 19th | 100362 | 20370 | 0 | 3268 | 0 | **124000** |
| 20th | 103921 | 21234 | 0 | 0 | 0 | **125155** |
| 21st | 43102 | 0 | 0 | 0 | 0 | **45310** |
| **Total** | **675114** | **236547** | **75096** | **7707** | **6183** | **1000647** |

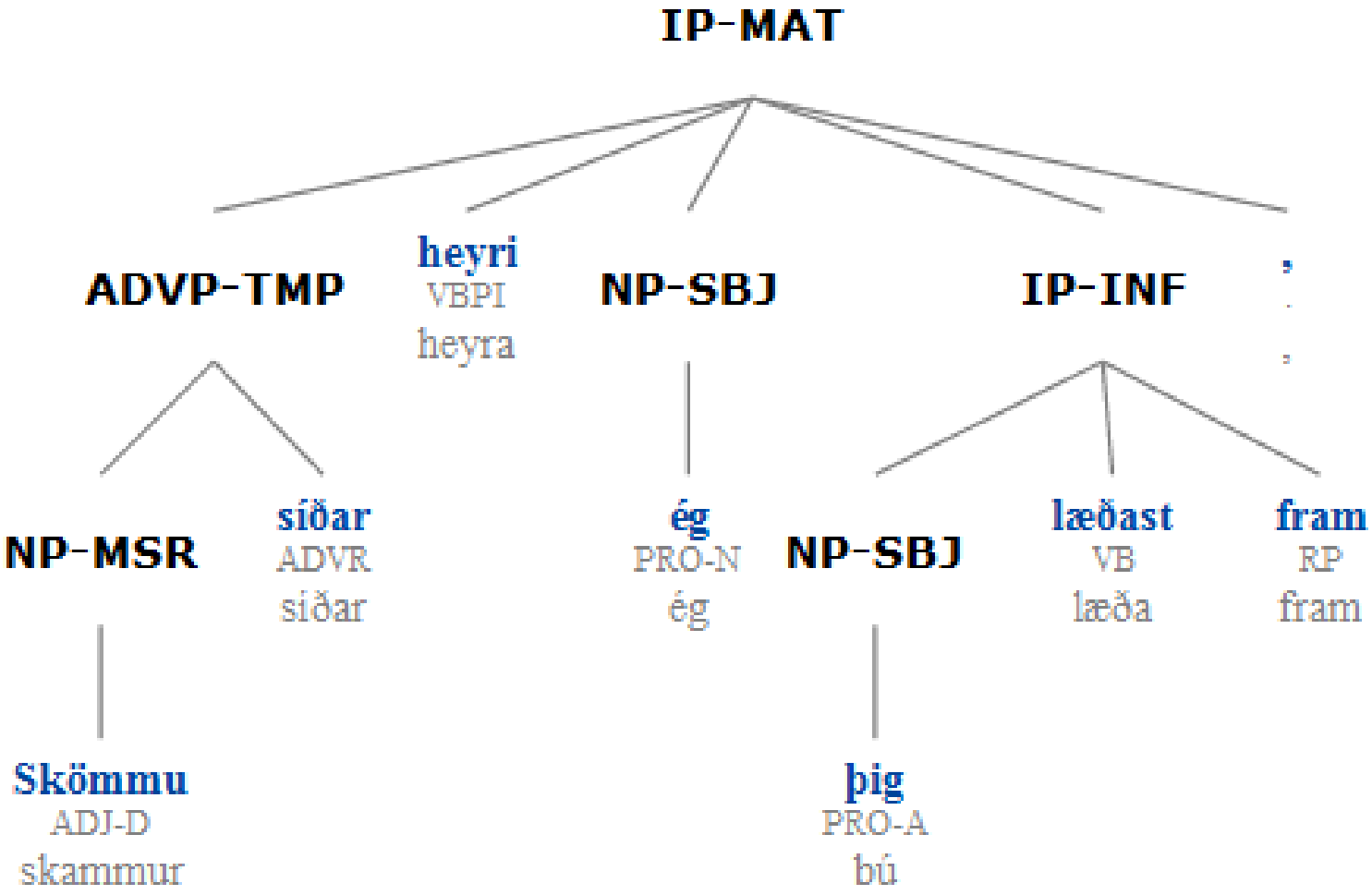
# Setningafræðileg þáttun

- Greining miðuð við sögulega enska trjábanka
  - Penn Corpora of Historical English
- Þó  löguð að íslensku eftir ástæðum og þörfum
  - nefnimynd (lemma) t.d. höfð með
  - fall nafnorða, fornafna og lýsingarorða einnig sýnt
- Ítarlegar greiningarleiðbeiningar eru til
  - fyrir sögulegu ensku trjábankana
  - en bætt var við þær eftir þörfum

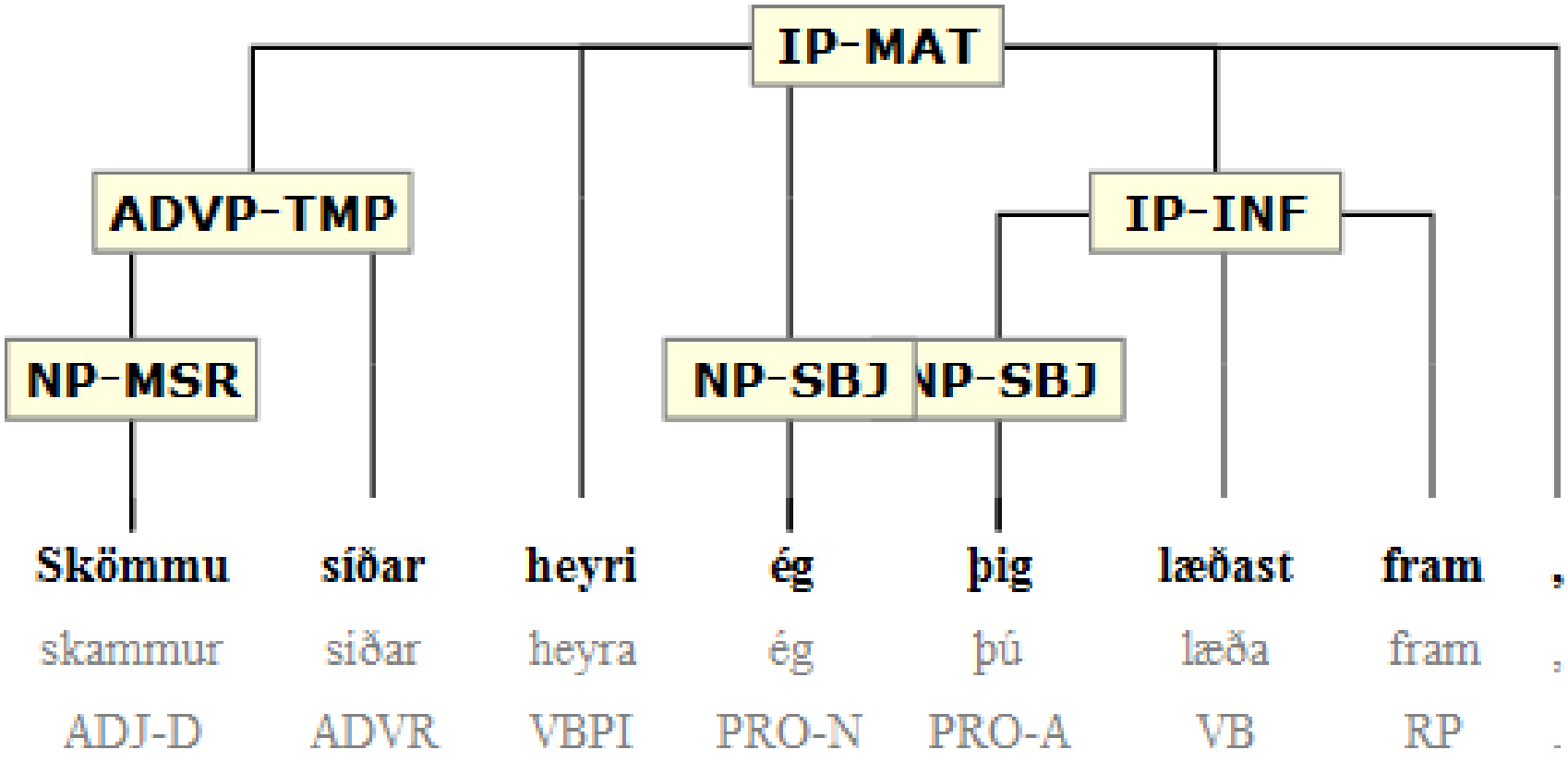
# Úttak úr IcePaHC

```
/~*
Skömmu-skammur síðar-síðar heyri-heyra ég-ég þig-þú læðast-læða fram-fram ,-,
(1985.SAGAN.NAR-FIC,.1295)
*~/
/*
1 IP-MAT:  14 NP-SBJ, 15 PRO-A
*/
( (IP-MAT (ADVP-TMP (NP-MSR (ADJ-D Skömmu-skammur))
                    (ADVR síðar-síðar))
          (VBPI heyri-heyra)
          (NP-SBJ (PRO-N ég-ég))
          (IP-INF (NP-SBJ (PRO-A þig-þú))
                  (VB læðast-læða)
                  (RP fram-fram))
          (. ,-,))
  (ID 1985.SAGAN.NAR-FIC,.1295))
```

# Í INESS er hægt að skoða setninguna svona



# ... og svona


IP-MAT
ADVP-TMP
IP-INF
NP-MSR
NP-SBJ
NP-SBJ
Skömmu síðar heyri ég þig læðast fram ,
skammur síðar heyra ég þú læða fram ,
ADJ-D ADVR VBPI PRO-N PRO-A VB RP .

# Leitað í trjábankanum

- Hægt er að sækja trjábankann í heild
  - ásamt leitarhugbúnaði fyrir Windows og Linux
    - http://www.linguist.is/icelandic_treebank/Download
- Leitað er með forritinu CorpusSearch
  - sem notar sérstakt fyrirspurnamál
  - og gerir kleift að setja fram flóknar fyrirspurnir
- Hér er leitað að þolfallsfrumlögum
  - NP-SBJ idoms *-A
    - idoms = 'immediately dominates' (beint yfirskipað)

# Leitarniðurstöður: NP-SBJ idoms *-A

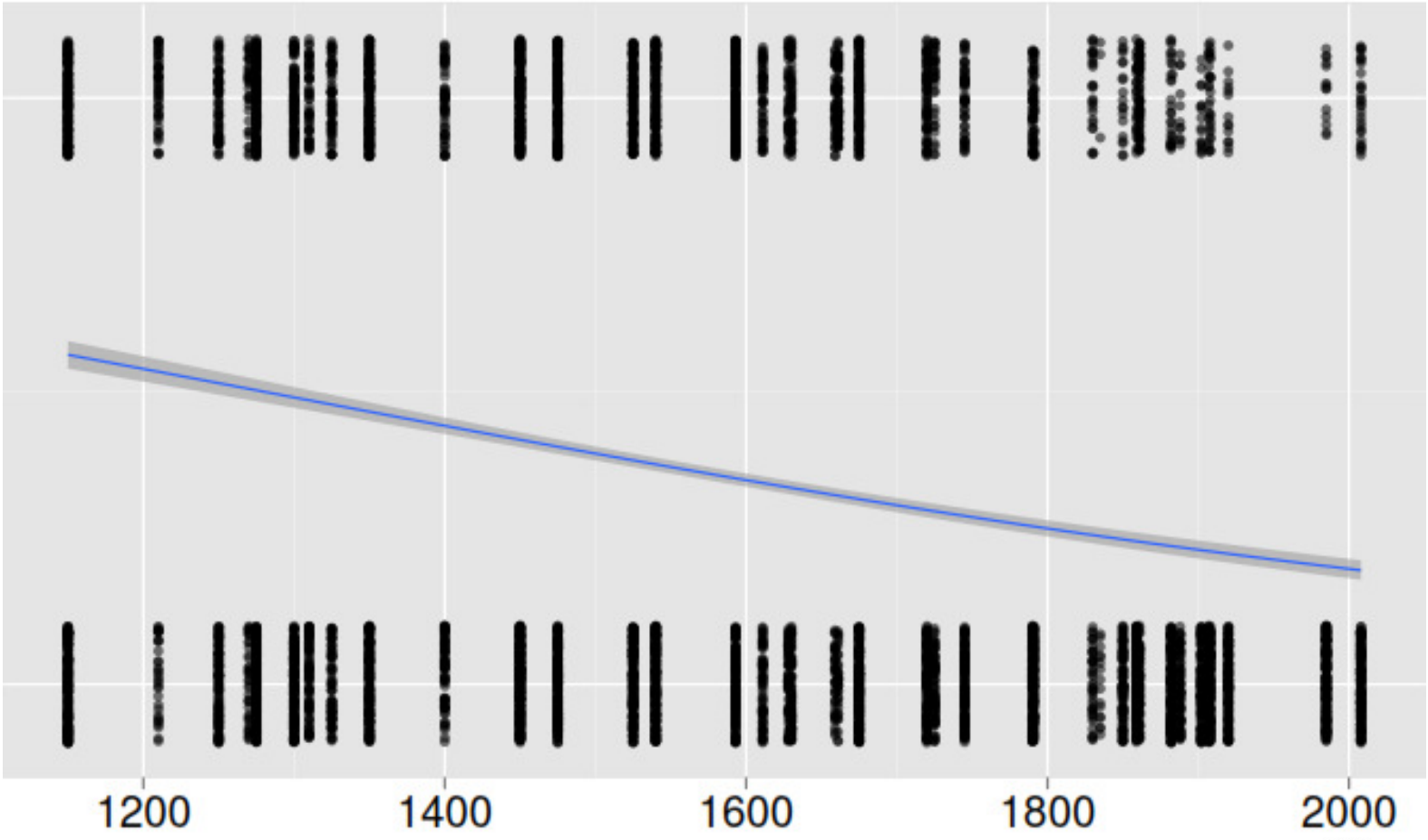
- SUMMARY:
- source files, hits/tokens/total
- 1150.firstgrammar.sci-lin.psd 13/13/182
- 1150.homiliubok.rel-ser.psd 123/123/2109
- 1210.jartein.rel-sag.psd 40/40/822
- 1210.thorlakur.rel-sag.psd 41/41/540
- 1250.sturlunga.nar-sag.psd 89/89/2224
- 1250.thetubrot.nar-sag.psd 17/17/246
- 1260.jomsvikingar.nar-sag.psd 84/84/1532
- 1270.gragas.law-law.psd 10/10/346
- 1275.morkin.nar-his.psd 90/90/2190
- 1300.alexander.nar-sag.psd 97/97/1383
- 1310.grettir.nar-sag.psd 99/99/2057
- 1325.arni.nar-sag.psd 69/69/1125
- 1350.bandamennM.nar-sag.psd 49/49/1223
- 1350.finnbogi.nar-sag.psd 132/132/2342
- 1350.marta.rel-sag.psd 62/62/977
- 1400.gunnar.nar-sag.psd 55/55/936
- 1400.gunnar2.nar-sag.psd 16/16/340
- 1400.viglundur.nar-sag.psd 51/51/1290
- 1450.bandamenn.nar-sag.psd 30/30/1034
- 1450.ectorssaga.nar-sag.psd 99/99/1950
- 1450.judit.rel-bib.psd 28/28/491
- 1450.vilhjalmur.nar-sag.psd 95/95/2414
- 1475.aevintyri.nar-rel.psd 70/70/1201
- 1480.jarlmann.nar-sag.psd 69/69/1283
- 1525.erasmus.nar-sag.psd 11/11/462
- 1525.georgius.nar-rel.psd 66/66/1060
- 1540.ntacts.rel-bib.psd 46/46/1185
- 1540.ntjohn.rel-bib.psd 41/41/1685
- 1593.eintal.rel-oth.psd 111/111/1552
- 1611.okur.rel-oth.psd 50/50/629
- 1628.olafuregils.bio-tra.psd 37/37/906
- 1630.gerhard.rel-oth.psd 43/43/701
- 1650.illugi.nar-sag.psd 113/113/1929
- 1659.pislarsaga.bio-aut.psd 30/30/324
- 1661.indiafari.bio-tra.psd 92/92/1388
- 1675.armann.nar-fic.psd 44/44/1018
- 1675.magnus.bio-oth.psd 23/23/204
- 1675.modars.nar-fic.psd 13/13/373
- 1680.skalholt.nar-rel.psd 19/19/870
- 1720.vidalin.rel-ser.psd 79/79/1112
- 1725.biskupasogur.nar-rel.psd 45/45/1105
- 1745.klim.nar-fic.psd 77/77/874
- 1790.fimmbraedra.nar-sag.psd 89/89/1603
- 1791.jonsteingrims.bio-aut.psd 50/50/1518
- 1830.hellismenn.nar-sag.psd 56/56/1411
- 1835.jonasedli.sci-nat.psd 10/10/163
- 1850.piltur.nar-fic.psd 64/64/1440
- 1859.hugvekjur.rel-ser.psd 87/87/1107
- 1861.orrusta.nar-fic.psd 32/32/1804
- 1882.torfhildur.nar-fic.psd 80/80/2000
- 1883.voggur.nar-fic.psd 4/4/130
- 1888.grimur.nar-fic.psd 12/12/625
- 1888.vordraumur.nar-fic.psd 44/44/759
- 1902.fossar.nar-fic.psd 74/74/1659
- 1907.leysing.nar-fic.psd 75/75/1521
- 1908.ofurefli.nar-fic.psd 63/63/1743
- 1920.arin.rel-ser.psd 88/88/1149
- 1985.margsaga.nar-fic.psd 65/65/1705
- 1985.sagan.nar-fic.psd 78/78/2008
- 2008.mamma.nar-fic.psd 76/76/1845
- 2008.ofsi.nar-sag.psd 53/53/1210
- whole search, hits/tokens/total 3568/3568/73014

# Unnið með leitarniðurstöður

- Þessar niðurstöður má færa inn í tölfræðiforrit
  - Excel, SPSS, R, o.s.frv.
- Síðan er hægt að skoða margvísleg atriði
  - bæði samtímalega og sögulega
  - athuga tíðni, reikna marktækni o.s.frv.
- Samanburður við önnur mál er líka auðveldur
  - ef greiningarskemað er hið sama
  - eins og í sögulegu ensku trjábönkunum o.fl.

# Dæmi: Þróun OV → VO í íslensku

- OV
- VO



# IcePaHC í INESS

- Bankinn hefur verið settur inn í INESS
  - safn trjábanka við Háskólann í Bergen
- Þar er hægt að skoða einstakar setningar
  - og leita að ákveðnum formgerðum
  - með svipuðu fyrirspurnamáli og í CorpusSearch
    - NP-SBJ > [pos="PRO-A"]
- Hins vegar fæst ekki tölfræði um leitina
  - fjöldi dæma úr hverjum texta kemur ekki fram

# Aðgengi og nýting

- Bankinn er opinn og ókeypis til niðurhals
  - ekki háður neinum leyfum
  - flestir textar löngu komnir úr höfundarrétti
  - en leyfi höfunda var fengið fyrir nýlegum textum
- Hver sem er má gera hvað sem er við bankann
  - við teljum að það sé allra hagur
  - bankinn hefur nú þegar verið nýttur í rannsóknum
  - og verður vonandi nýttur í máltækni á næstunni

# Færeyskur trjábanki – og fleiri?

- Nú er hafin gerð færeysks trjábanka, FarPaHC
  - verkefni á vegum Eiríks Rögnvaldssonar
    - styrkt af Rannsóknasjóði Háskóla Íslands
  - sem Einar Freyr Sigurðsson vinnur að
- Færeyski trjábankinn verður sniðinn að IcePaHC
  - reynt að afla einhverra sömu texta
    - a.m.k. úr Nýja testamentinu
- Sótt hefur verið um styrk til víðtækara samstarfs
  - um hliðstæða trjábanka fyrir norræn mál o.fl.

# Tilvísanir


- Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. http://www.linguist.is/icelandic_treebank
- Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson og Joel Wallenberg. 2011. Creating a Dual-Purpose Treebank. *Journal for Language Technology and Computational Linguistics* 26,2:141-152.