Þáttun og trjábankar

Hrafn Loftsson

Eiríkur Rögnvaldsson

Anton Karl Ingason

Hugvísindaþing

14. mars 2009

Þáttun (e. parsing)

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

IceParser

- Hlutaþáttari fyrir íslenskan texta
  - (Hrafn Loftsson og Eiríkur Rögnvaldsson, 2007)
- Inntak: Markaður texti
- Úttak: Þáttaður texti
  - setningarliðir merktir
  - setningafræðileg hlutverk merkt
- Útfærður með svokölluðum stöðuferjöldum
- Mjög hraðvirkur
  - hægt að prófa á http://nlp.ru.is
    - „Augnaráðið negldist við gráa jakkann sem hann var að klæða sig úr og hengja inn í skáp.“

Dæmi um greiningu

{*SUBJ> [NP augnaráðið nheng NP] ***SUBJ>}**
[VP negldist sfm3eþ VP]
[PP við ao [NP [AP gráa lkeovf AP] jakkann nkeog NP] PP]
[CP sem ct CP]
{*SUBJ> [NP hann fþken NP] ***SUBJ>}**
[VPb var sfg3eþ VPb]
[VPi að cn klæða sng VPi]
{*OBJ< [NP sig fþkeo 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]

Útfærsla

- Stigvaxandi þáttari byggður á stöðuferjöldum
  - e. incremental finite-state parser
- Sérhvert stöðuferjald:
  - hefur það hlutverk að bera kennsl á tiltekið mynstur í inntaki
  - skrifar greiningarupplýsingar inn í inntakstextann
  - skilar breyttum texta út, tilbúnum til meðhöndlunar fyrir næsta stöðuferjald

Stöðuferjöld: Setningarliðir

- Keyrð í tiltekinni röð – einfaldir liðir fyrst
- Fyrst eru atviksliðir merktir
  - ... var sfg3eþ [AdvP mjög aa AdvP] gott lhensf félagslíf nhen
- Síðan lýsingarorðsliðir
  - ... var sfg3eþ [AP [AdvP mjög aa AdvP] gott lhensf AP] félagslíf nhen
- Síðan nafnliðir
  - ... var sfg3eþ [NP [AP [AdvP mjög aa AdvP] gott lhensf AP] félagslíf nhen NP]

Stöðuferjöld: Setningarliðir

- Síðan sagnliðir
  - [VPb var sfg3eþ VPb] [NP [AP [AdvP mjög aa AdvP] gott lhensf AP] félagslíf nhen NP]
- Síðan forsetningarliðir
  - [PP af aþ [NP þessum fakfþ [AP stöðugu lkfþvf AP] ósigrum nkfþ mínum fekfþ NP] PP]
- Stöðuferjöld fyrir setningafræðileg hlutverk vinna á sambærilegan hátt og stöðuferjöld fyrir setningarliði
  - eitt ferjald sér um að merkja frumlög, annað andlög, hið þriðja sagnfyllingar, o.s.frv.

Trjábankar

- Trjábanki (treebank)
  - málheild með setningafræðilegri greiningu
- Trjábankar eru af ýmsu tagi
  - sumir byggjast á ákveðnu kenningakerfi
    - t.d. HPSG
  - aðrir leitast við að vera óháðir kenningakerfum
- Elstur og þekktastur er Penn Treebank
  - gerður við University of Pennsylvania

Gagnsemi og gerð trjábanka

- Trjábankar eru mjög mikilvæg tól
  - í setningafræðilegum rannsóknum af ýmsu tagi
  - og ekki síður í máltækni
- Trjábankar fyrir mörg tungumál eru í smíðum
  - en gerð þeirra er mjög tímafrek og dýr
- Hvernig fáum við trjábanka fyrir íslensku?
  - hefðbundnar aðferðir alltof dýrar
  - nauðsynlegt að þróa hagkvæmari leiðir

Tilraun í UPenn

- IceParser keyrður á markaðan texta
  - úr *Íslenskri orðtíðnibók*
- Íslenskum mörkum breytt í ensk
  - PPCEME sem notuð eru í Early Modern English
- Útkoman þáttuð með þáttara þjálfuðum á EME
  - Collins/Bikel parser
- Útkoman leiðrétt í höndunum
  - sem reyndist ekki vera ýkja mikið verk

Úttak úr þáttara

```
(IP-MAT (NP-SBJ (NS viðskiptafræðinemar)) (VBD
    ætluðu)
  (IP-INF (TO að)
          (VB halda)
          (NP-MSR (Q+WPRO eitthvað))
        (CP-ADV (C sem)
                (IP-SUB (NP-SBJ (PRO þeir))
                  (VBD kölluðu)
                    (NP-OB1 (N pabbakvöld)))))
(. .))
```

Eftir handvirka leiðréttingu

```
(IP-MAT (NP-SBJ (NS viðskiptafræðinemar)) (VBD
    ætluðu)
  (IP-INF (TO að)
          (VB halda)
          (NP-ACC (Q+WPRO eitthvað))
        (CP-REL (C sem)
                        (IP-SUB (NP-SBJ (PRO þeir))
                          (VBD kölluðu)
                                (NP-OB1 (N pabbakvöld)))))))
(. .))
```

Þáttun án orða

- Hvernig er þetta hægt?
  - að láta þáttara þjálfaðan á ensku þátta íslensku?
- Þáttun byggist venjulega m.a. á orðasafni
  - því þarf að byrja með þjálfunarsafn
  - og þarf að þjálfa þáttara á hverju máli fyrir sig
- Hér byggist þáttun eingöngu á málkerfi
  - þáttarinn miðar bara við mörk og orðaröð
  - en lítur alveg fram hjá orðunum

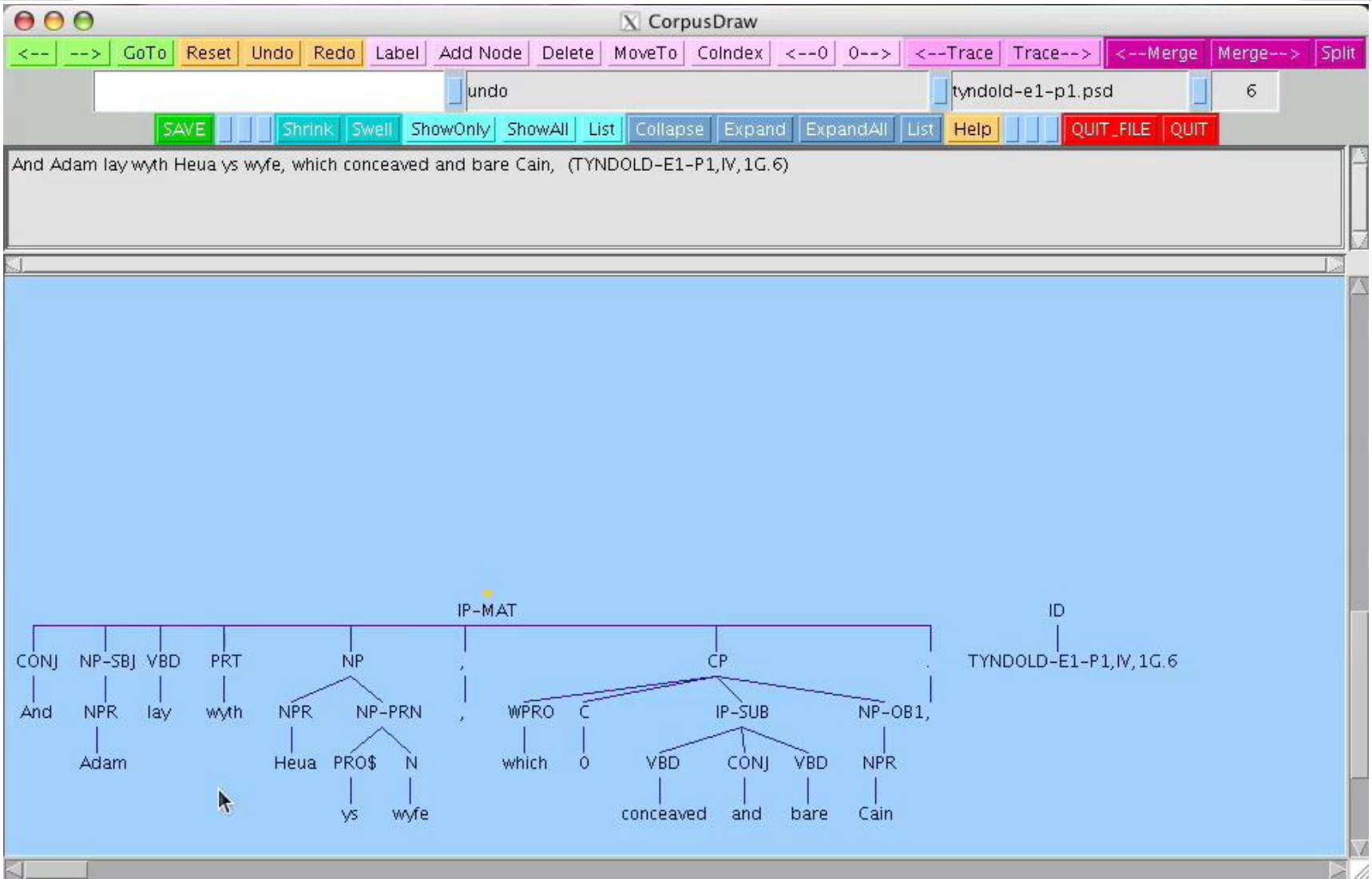
Samspil málfræði og tölfræði

- Byggt verður á þessari tilraun
  - en reynt að koma meiri málfræði í greininguna
- Ætlunin er að greina beygingarlega þætti
  - sem tölfræðiþáttari getur nýtt sér
  - og bæta þeim inn í inntak hans
- Þetta má t.d. gera með breytingu á IceParser
  - láta hann greina og skrifa út fleiri þætti

Handvirk leiðrétting

- Handvirk leiðrétting er mjög tímafrek
  - því skiptir máli að þróa aðferðir til að flýta henni
- Við munum nota hugbúnaðinn CorpusDraw
  - sem skrifaður er við UPenn
- Ætlunin er að endurbæta hugbúnaðinn
  - Anton Karl Ingason mun vinna að því
  - fer til UPenn í næstu viku til að undirbúa það

Skjámynd úr CorpusDraw



Gerð íslensks trjábanka

- Að þessu loknu hefst smíði íslensks trjábanka
  - þar sem byrjað verður á vélrænni þáttun
  - með eftirfarandi handvirkri leiðréttingu
- Síðan verður þáttarinn þjálfaður á ný
  - á leiðrétta textanum
- Því næst er þáttarinn látinn greina meiri texta
  - sem verður einnig leiðréttur í höndunum
  - og svo koll af kolli

Greiningartextar

- Hversu mikinn texta á að greina?
  - ræðst af því hversu tímafrek greiningin verður
  - samstarfsaðilar hafa talað um 1,5 milljónir orða
- Hvers konar texta á að greina?
  - dæmigerða texta úr nútímamáli
  - væntanlega úrval úr markaðri íslenskri málheild
  - en líka eldri texta frá ýmsum tímum málsögunnar
  - svo að trjábankinn nýtist til að skoða málbreytingar

Fyrirmynd

- Með þessu gerum við ráð fyrir að fá trjábanka
  - fyrir brot af þeim kostnaði sem venjulegur er
  - en ekki víst að allt verði leiðrétt í höndunum
- Aðferðafræðinni er ætlað að vera fyrirmynd
  - sem nýst geti öðrum tungumálum
  - þar sem litlar mállegar gagnalindir eru til
  - og/eða beygingar eru miklar


Þökk fyrir áheyrnina

hrafn@ru.is
eirikur@hi.is
anton.karl.ingason@gmail.com