Hlutaþáttun íslensks texta

Verkefni til eins árs

Styrkt af Rannís

3,8 m.kr.

Samvinna og samstarf

- Samvinna þriggja fræðigreina
  - málfræði: Eiríkur Rögnvaldsson, prófessor
  - tölvunarfræði: Hrafn Loftsson, MSc
  - tölfræði: Sigrún Helgadóttir, MSc
- Samstarf þriggja stofnana
  - Háskóli Íslands (Eiríkur)
  - Háskólinn í Reykjavík (Hrafn)
  - Orðabók Háskólans (Sigrún)

Setningafræðileg þáttun (parsing)

- 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

Mismunandi þáttun setningar

- Full þáttun – mismunandi greiningar:
  - $[_{S}\ [_{NL}$ Margir$]\ [_{SL}$ hittu $[_{NL}$ Maríu $[_{FL}$ á $[_{NL}$ skrifstofunni$]]]]]$
  - $[_{S}\ [_{NL}$ 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

Hvað er greint?

- [Gerð liða]
  - NP - nafnliður
  - APx - lýsingarorðsliður
  - AdvP - atviksliður
  - PP - forsetningarliður
  - CP - tengiliður
  - VPx - sagnliður
  - MWEx - orðasamband
- {*Hlutverk}
  - SUBJ - frumlag
  - OBJ - andlag
  - COMP - sagnfylling
  - QUAL - eignarfallseink.
  - X > - tengist so. á eftir
  - X < - tengist so. á undan

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

- “Incremental finite-state parser”
- Stigvaxandi þáttari byggður á endanlegum stöðuaðferðum
  - Röð af stöðuferjöldum (e. finite-state transducers).
  - 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

Stöðuferjöldin

- Skiptast í tvo flokka:
- Ferjöld sem greina setningarliði
  - atviksliði, lýsingarorðsliði, nafnliði, forsetningarliði, sagnliði, o.s.frv.
- Ferjöld sem greina setningafræðileg hlutverk
  - frumlög, andlög, sagnfyllingar, eignarfallseinkunnir

Stöðuferjöld: Setningarliðir

- Hönnunarforsendur:
  - Reynt að nýta beygingarleg einkenni sem minnst þegar setningarliðir eru greindir
  - Orðflokkur og röð orða látin stýra greiningu
  - Í stað þess að láta t.d. samræmi í kyni, tölu og falli stýra greiningu á nafnliðum
  - Af hverju?
    - Til að setningagreiningin nýtist betur fyrir málfræðileiðréttingu

Stöðuferjöld: Setningarliðir

- Keyrð í tiltekinni röð – þeir einföldu 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: Setningarliðir

- Fleiri ferjöld:
  - Merkja fleiryrt orð/orðasambönd, eins og:
    - [PP **[MWE_PP út aa um ao MWE_PP]** [NP gluggann nkeog NP] PP]
    - **[MWE_AdvP allt fohen í aþ einu lheþsf MWE_AdvP]**
    - [PP í aþ [NP **[MWE_AP neins fokee konar nkee MWE_AP]** samfloti nheþ NP] PP]
  - Merkja haus (aðalorð) nafn- og lýsingarorðsliða
    - [NPn ég fp1en ***HeadNn** NP]
    - [NPa stólinn nkeog ***HeadNa** NP]

Stöðuferjöld: Setningafræðileg hlutverk

- Ferjöldin nýta sér setningarliðamerkingar og hausamerkingar undanfarandi ferjalda
- Sérstakt stöðuferjald merkir eignarfallseinkunnir
  - [NPa [APa síðustu **lveove** *HeadAa AP] nóttina **nveog** *HeadNa NP] {****QUAL** [NPg **okkar** **fp1fe** ***HeadNg** NP] ***QUAL}**
- Sérstakt stöðuferjald merkir frumlög
  - **{*SUBJ>** [NPn **ég** **fp1en** ***HeadNn** NP] ***SUBJ>}** [VP tók **sfg1eþ** VP] [NPa ákvörðun **nveo** *HeadNa NP]
  - [NPa hvað **fsheo** *HeadNa NP] [VP á **sfg1en** VP] **{*SUBJ<** [NPn **ég** **fp1en** ***HeadNn** NP] ***SUBJ<}** [VPi að **cn** segja **sng** VPi] ? **?**

Stöðuferjöld: Setningafræðileg hlutverk

- Sérstakt stöðuferjald merkir **andlög** og **sagnfyllingar**
- Nýtir sér setningarliðamerkingar, frumlags-merkingar og upplýsingar um föll
  - **{*SUBJ>** [NPn hún fpven *HeadNn NP] ***SUBJ>}** [VP sæi svg3eþ VP] **{*OBJ<** [NPa **mig** fp1eo ***HeadNa** NP] ***OBJ<}**
  - **{*SUBJ>** [NPn ég fp1en *HeadNn NP] ***SUBJ>}** [VP veitti sfg1eþ VP] **{*IOBJ<** [NPd **því** fpheþ ***HeadNd** NP] ***IOBJ<}** **{*OBJ<** [NPa **athygli** nveo ***HeadNa** NP] ***OBJ<}**
  - **{*SUBJ>** [NPn ég fp1en *HeadNn NP] ***SUBJ>}** [VPb er sfg1en VPb] **{*COMP<** [APn **viss** lkensf ***HeadAn** AP] ***COMP<}**

Að lokum: “Hreingerningarferjöld”

- Ferjöld sem snyrta til textann
  - Eyða aukabilum
  - Eyða sérstökum merkingum sem eru notuð til hjálpar
  - Skrifa einn setningarlið í hverja línu

{*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 fpken NP] ***SUBJ>}**
[VPb var sfg3eþ VPb]
[VPi að cn klæða sng VPi]
{*OBJ< [NP sig fpkeo 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]