Keeletehnoloogia seminar

suulise keele süntaktilisest analüüsist

Kaili Müürisep
kaili.muurisep@ut.ee

Eesti keele süntaktilisest analüüsist

- On olemas kirjaliku keele morfosüntaktiline ühestaja ja süntaksianalüsaator, mis põhineb kitsenduste grammatikal
- Grammatika koosneb 1300+1200 reeglist
- Annab madala pindmise süntaktilise kirjelduse 85-90% sõnadest korrektsusega 96-98%

Näide analüüsi käigust

- Eesti vanimad asukad saabusid siia pärast viimast jääaega.

Morfoloogiliselt analüüsitud lause

```
Eesti
    Eesti+0 //_S_ prop sg gen #cap //
    Eesti+0 //_S_ prop sg nom #cap //
    eesti+0 //_G_ #cap //
vanimad
    vanim+d //_A_ super pl nom  //
asukad
    asukas+d //_S_ com pl nom  //
saabusid
    saabu+sid //_V_ main indic impf ps2 sg ps af  #Intr //
    saabu+sid //_V_ main indic impf ps3 pl ps af  #Intr //
siia
    siia+0 //_D_ //
    siig+0 //_S_ com sg gen  //
```

```
pärast
    pärast+0 //_D_ //
    pärast+0 //_K_ post #gen //
    pärast+0 //_K_ pre #part //
    pärane+t //_A_ pos sg part  //
    pära+st //_S_ com sg el  //
viimast
    viimane+t //_A_ pos sg part  //
    vii+mast //_V_ main sup ps el  #NGP-P //
jääaega
    jää_aeg+0 //_S_ com sg adit //
    jää_aeg+0 //_S_ com sg part //
$.
    . //_Z_ Fst //
```

Morfoloogiliselt ühestatud näitelause

Eesti
    Eesti+0 //_S_ prop sg gen #cap //
vanimad
    vanim+d //_A_ sup pl nom //
asukad
    asukas+d //_S_ com pl nom //
saabusid
    saabu+sid //_V_ main indic impf ps3 pl ps af #Intr //
siia
    siia+0 //_D_ //
pärast
    pärast+0 //_K_ pre #part //
viimast
    viimane+t //_A_ pos sg part //
jääaega
    jää_aeg+0 //_S_ com sg part //

Pärast süntaktiliste märgendite lisamist

Eesti
    Eesti+0 //_S_ prop sg gen #cap //**CLB @OBJ @ADVL @NN> @<NN @<Q
vanimad
    vanim+d //_A_ super pl nom // @ADVL @AN> @<AN @PRD
asukad
    asukas+d //_S_ com pl nom //@SUBJ @PRD @OBJ @NN> @<NN @ADVL @<Q
saabusid
    saabu+sid //_V_ main indic impf ps3 pl ps af #Intr // @+FMV
siia
    siia+0 //_D_ // @ADVL @AD> @<AD
pärast
    pärast+0 //_K_ pre #part // @ADVL @PN> @<PN
viimast
    viimane+t //_A_ pos sg part // @AN> @<AN @ADVL
jääaega
    jää_aeg+0 //_S_ com sg part // @SUBJ @OBJ @ADVL @<Q @NN> @<NN @<P

Süntaktiliselt analüüsitud näitelause

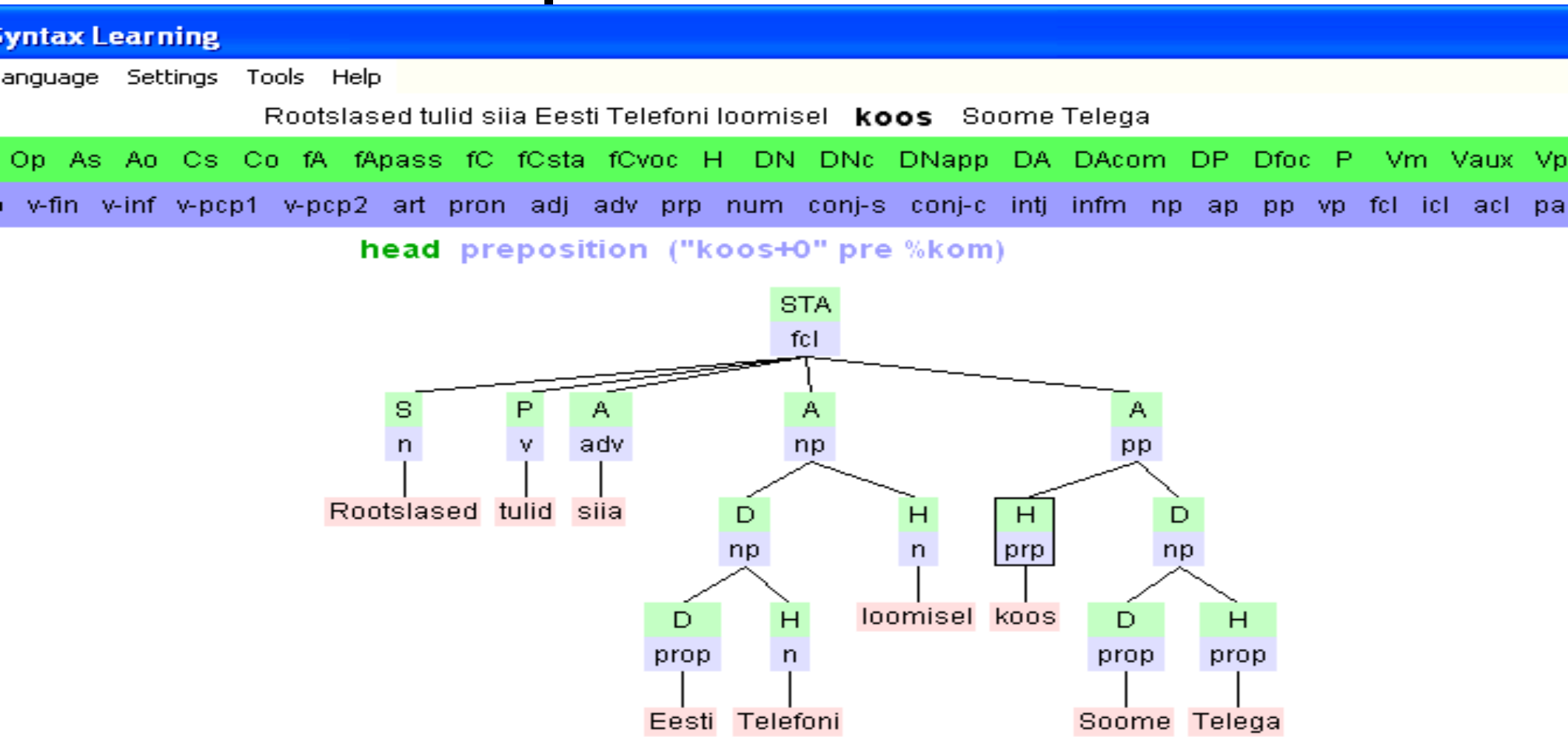
```
Eesti
    Eesti+0 //_S_ prop sg gen #cap // **CLB  @NN>
vanimad
    vanim+d //_A_ super pl nom  //  @AN>
asukad
    asukas+d //_S_ com pl nom  //  @SUBJ
saabusid
    saabu+sid //_V_ main indic impf ps3 pl ps af  #Intr //  @+FMV
siia
    siia+0 //_D_ //  @ADVL
pärast
    pärast+0 //_K_ pre #part //  @ADVL
viimast
    viimane+t //_A_ pos sg part  //  @AN>
jääaega
    jää_aeg+0 //_S_ com sg part //  @<P
```

Näide reeglitest

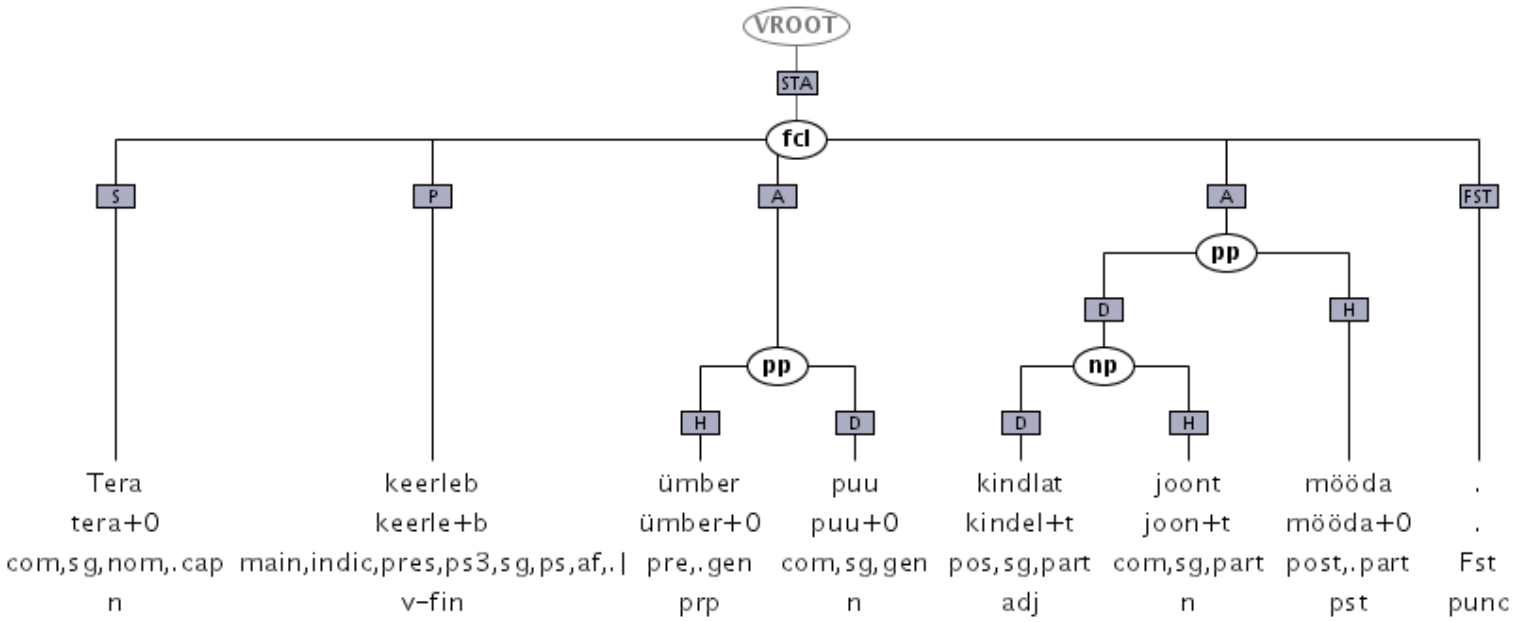
(@w=s! (@AN>) (0 SgNom)(1 Subst)(1 SgNom))

(@w=s0 (@OBJ) (*-1 Prep *R+1)(NOT *R+1 PrepComp)
(*1 PrepComp *L-1)(NOT *L-1 Prep) **CLB)

Puudepank – Arborest VISLis



Puudepank – Rätsepa laused TIGER XMLis



Suulise keele süntaktiliselt märgendatud korpus

- 100000 sõna morfoloogiliselt ühestatud teksti
- morfoloogiline info käsitsi parandatud (kolmkend = kolmkümmend)
- täiesti käsitsi morfoloogiliselt ühestatud
- osa transkriptsioonimärgendeid eemaldatud

Näide tekstist

```
$<s> ???K    ####
    ####
ja
    ja+0 //_J_ crd //  **CLB @J
on
    ole+0 //_V_ main indic pres ps3 sg ps af #FinV #Intr //  @+FMV
aeroobikatreener
    aeroobika_treener+0 //_S_ com sg nom //  @SUBJ
seal
    seal+0 //_D_ //  @ADVL
$.
    . //_Z_ Fst //
väliseestlastele
    välis_eestlane+tele //_S_ com pl all //  **CLB @ADVL
$.
    . //_Z_ Fst //
aasta
    aasta+0 //_S_ com sg nom //  **CLB @P>
otsa
    otsa+0 //_K_ post #nom //  @ADVL
$.
    . //_Z_ Fst //
```

Täiendused

- Süntaktilised märgendid
  - @B partikkel *öö*, *noh*, *mhmh*
  - @T tundmatu sõna või tundmatu süntaktiline funktsioon
    - teeme ee need natuke harjutusi nende peale võibolla
  - Minimaalselt muudeti süntaktilisi kitsendusi
  - Täiesti uued osalausepiirid

Tulemused (NODALIDA-05)

- the word count in the corpus: 2543
- recall (the ratio of the number of correctly assigned syntactic tags to the number of all correct tags): 97.3% (98.5%)
- precision (the ratio of the number of correctly assigned syntactic tags to the number of all assigned syntactic tags) : 89.2% (87.5%)
- unambiguity rate: 91.5% (89.5%)

Vead

- osalausepiirid ca 17%
  - selle taga on saad aru selline lähenemine
- tundmatu süntaktiline funktsioon 19%
- adjektiivid käituvad nimisõnana 12%
- üks viga põhjustab teise 5%
- kordused 3%
  - noh se see on tähtis

Mitteladususued

- Korduste märgendamine
  - miks miks miks peab ...
- Paranduste märgendamine
  - väga nor- väga normaalne noh väga naiss
- Pealerääkimised
  - et nad
    mhmh
    sobivad kätte

Organisatsioonilised küsimused

- Alustaks 10.30?
- Rahvaloendus
- 2 ettekannet ühel päeval?
- Millest?

Teemad

- Puudepangad
  - Treebanks for Spoken Language – some reflections (J. Allwood)
  - Corpus of Spoken Swedish
  - VERBMOBIL Treebank
  - Spoken Dutch Corpus
- Suulise keele teksti madalsüntaktiline märgendamine
  - Annotating and parsing spoken language (Johannessen & Jorgensen)
  - Morfosüntaktiline ühestamine
- Mis on lausung
- Mitteladusused
- Paranduste automaatne tuvastamine