Puudepangad

Kaili Müürisep
25.01.07

Ülevaade

- Ülevaade puudepankade liikidest
- Eestikeelsed süntaktiliselt märgendatud korpused
- VISL Tree Editor
- Metsast

Ajaloost

- Puudepank – süntaktiliselt märgendatud korpus
  - moodustajate struktuur
  - funktsionaalne struktuur
  - kompromiss
- Esimesed tekkisid 30 a tagasi
- Tuntuim Penn Treebank

Fraasistruktuuriga puudepank

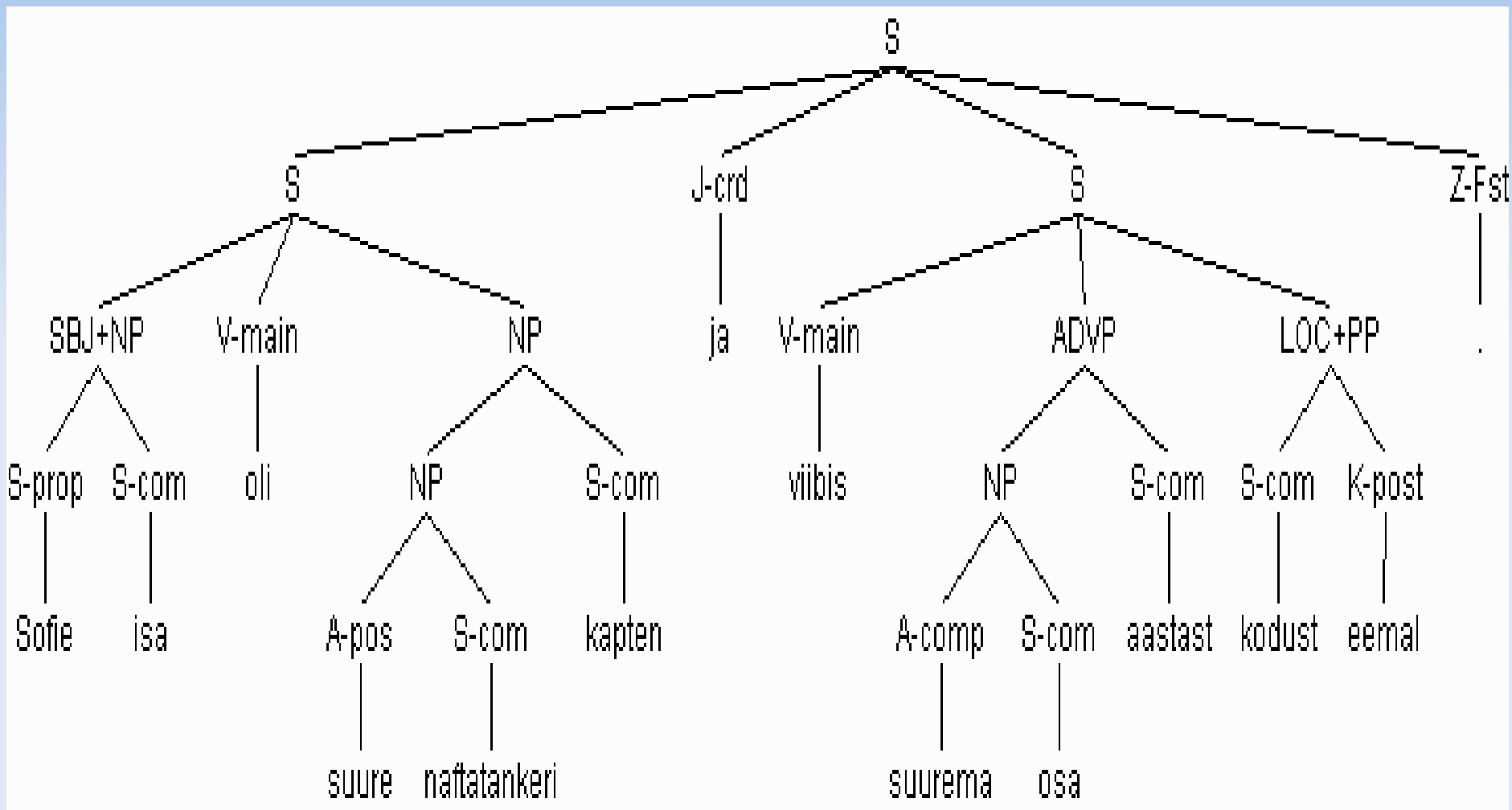
- Tuntuim esitusviis on suluesitus:
  - Lancaster Parsed Corpus

[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1 Pier_NNL1 N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RR [P at_II [N his_APP$ new_JJ home_NN1 [P in_II [N Windsor_NP1 [safari_NN1 park_NNL1]N]P]N]P]V] ._. S]

- Penn Treebank (ajalooline)

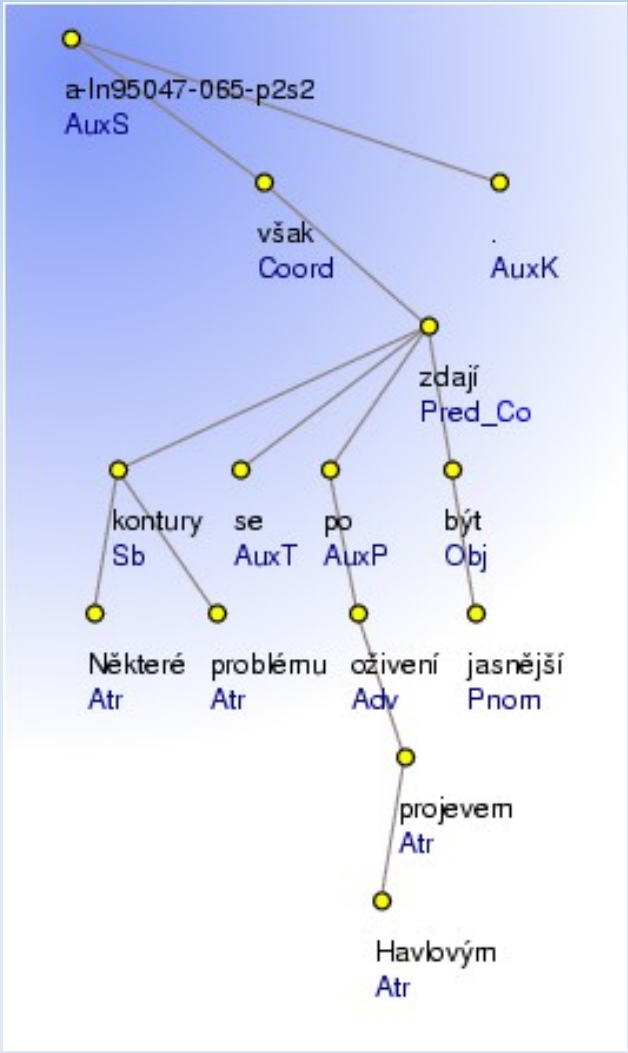
```
( (S
      (NP Martin Marietta Corp.)
      was
      (VP given
            (NP a  $ 29.9  million Air Force contract
                  (PP for
                        (NP low-altitude navigation and targeting equipment)))))
.)
```

Puuesitus


S
S
J-crd
S
Z-Fst
SBJ+NP
V-main
NP
ja
V-main
ADVP
LOC+PP
.
S-prop
S-com
oli
NP
S-com
viibis
NP
S-com
S-com
K-post
Sofie
isa
A-pos
S-com
kapten
A-comp
S-com
aastast
kodust
eemal
suure
naftatankeri
suurema
osa

Sõltuvuspuude pangad

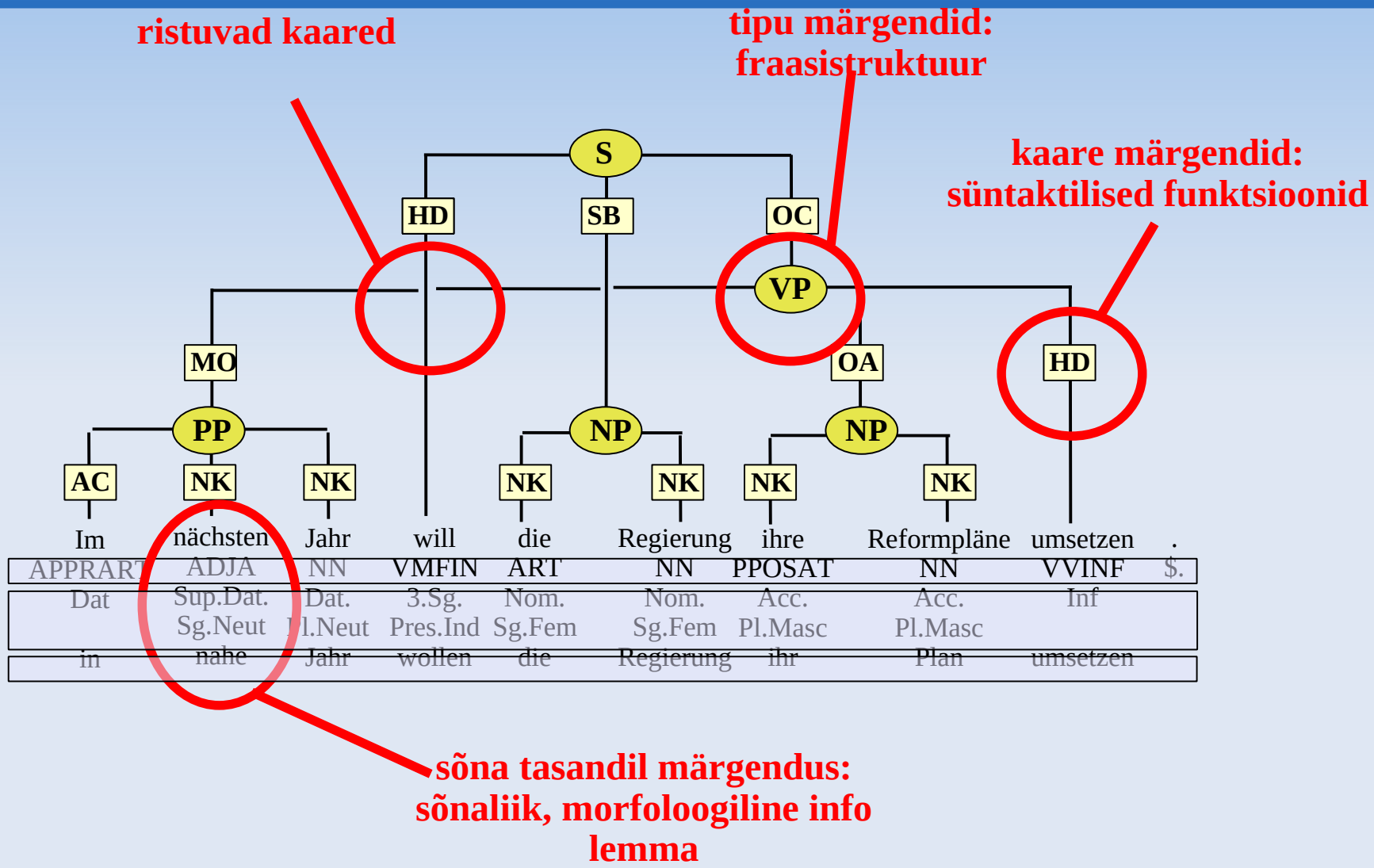
- Suurim Praha puudepank – 90000 lauset (2005?)



TIGER Treebank

- TIGER saksa keele puudepank – 50000 lauset, koos NEGRA korpuse 20000 lausega võrreldav Penn Treebankiga (1.5 milj sõna)
- TIGER: LinguisTic Interpretation of a GERman Corpus

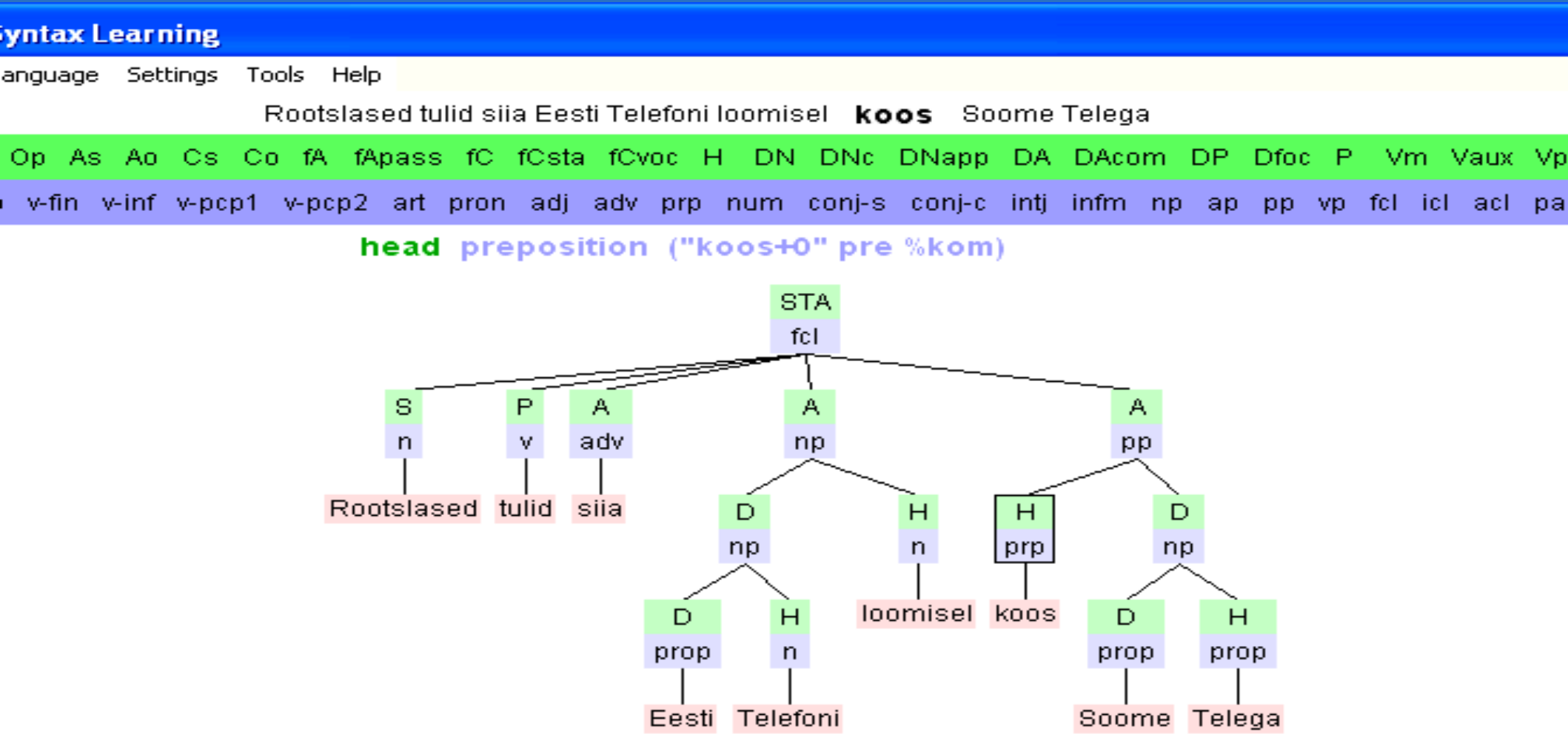
TiGer: märgendus



VISL treebanks

- Arboretum (Danish treebanks)
- Floresta sintá[c]tica (Portuguese treebanks)
- L'Arboratoire (French treebanks)
- Arborest (Estonian treebanks)

Arborest


yntax Learning
anguage Settings Tools Help
Rootslased tulid siia Eesti Telefoni loomisel koos Soome Telega
Op As Ao Cs Co fA fApass fC fCsta fCvoc H DN DNc DNapp DA DAcom DP Dfoc P Vm Vaux Vp
v-fin v-inf v-pcp1 v-pcp2 art pron adj adv prp num conj-s conj-c intj infm np ap pp vp fcl icl acl pa
head preposition ("koos+0" pre %kom)
STA
fcl
S
n
Rootslased
P
v
tulid
A
adv
siia
A
np
D
np
D
prop
Eesti
H
n
Telefoni
H
n
loomisel
A
pp
H
prp
koos
D
np
D
prop
Soome
H
prop
Telega

Arborest puu formaat

```
SOURCE id='est-125'
ID='est-125' Rootslased tulid siia Eesti Telefoni loomisel koos
   Soome Telega.
A1/2
STA:fcl
S:n("rootslane+d" com pl nom .cap)   Rootslased
P:v("tule+id" main indic impf ps3 pl ps af .FinV .Intr)   tulid
A:adv("siia+0")siia
A:np
=D:np
==D:prop("Eesti+0" prop sg gen .cap)    Eesti
==H:n("telefon+0" com sg gen .cap)  Telefoni
=H:n("loomis+l" com sg ad)loomisel
A:pp
=H:prp("koos+0" pre .kom) koos
=D:np
==D:prop("Soome+0" prop sg gen .cap) Soome
==H:prop("Tele+ga" prop sg kom .cap)   Telega
```

Eestikeelsed süntaktiliselt märgendatud korpused

- EstCG korpus – pindmine sõltuvusmärgendus, 350000 sõna
- Sofie paralleelpuudepank – fraasistruktuurimärgendus, 50 lauset
- Arborest – VISL-märgendusega puudepank, 150 lauset

EstCG korpus

<s>
Mitmekesisus
    mitme_kesi=sus+0 //_S_ com sg nom #cap // **CLB @SUBJ
on
    ole+0 //_V_ main indic pres ps3 sg ps af #FinV #Intr // @+FMV
elu
    elu+0 //_S_ com sg gen // @NN>
vaieldamatu
    vaieldamatu+0 //_A_ pos sg nom // @AN>
omapära
    oma_pära+0 //_S_ com sg nom // @PRD
$.
    . //_Z_ Fst //
</s>

EstCG2

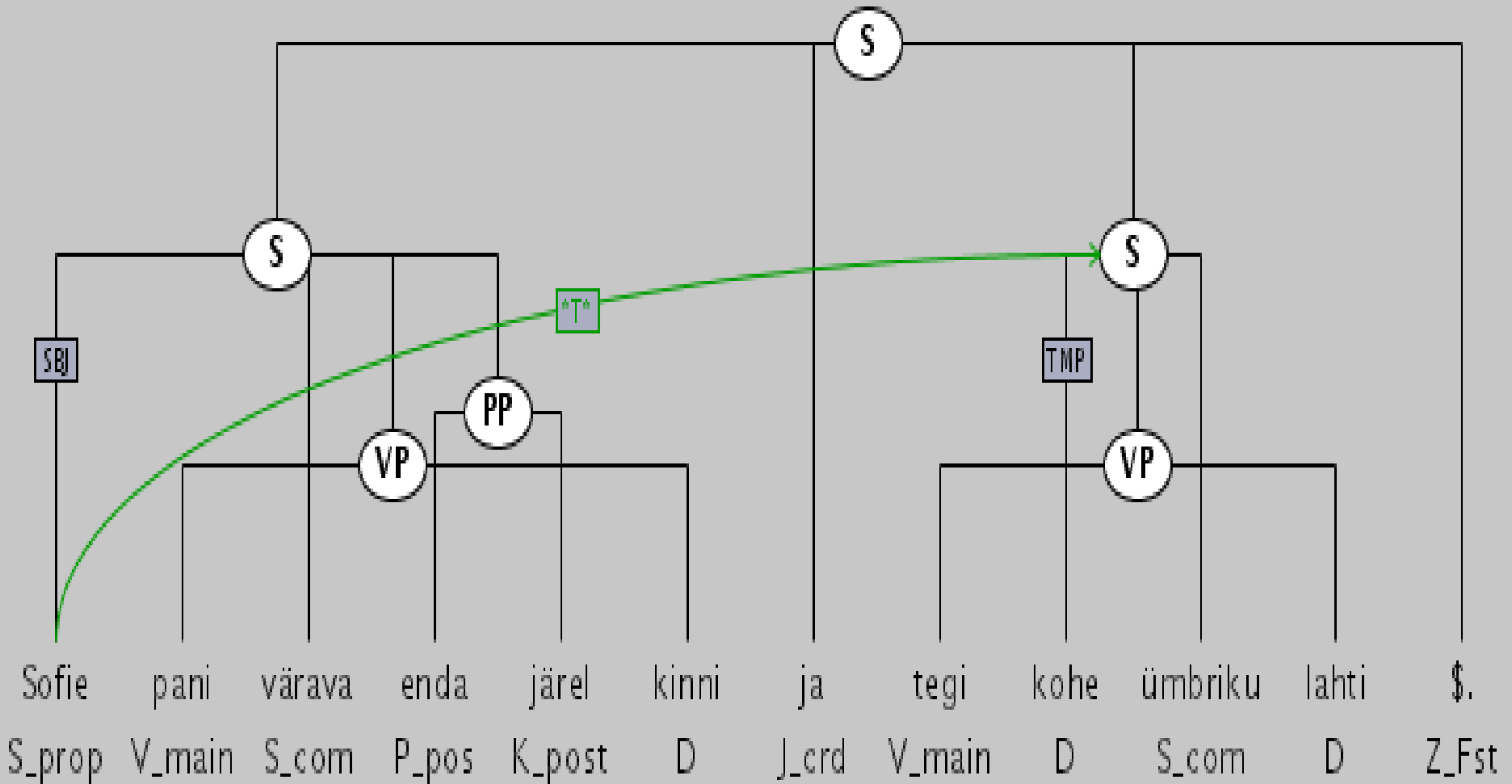
```
"<Samas>"
        "samas" <samas+0> D cap @ADVL
"<mõistis>"
        "mõist" <mõist+is> V main indic impf ps3 sg ps af FinV #NGP-P InfP @FMV
"<Rasmus>"
        "Rasmus" <Rasmus+0> S prop sg nom cap @SUBJ
"<jahmatusega>"
        "jahmatus" <jahmatus+ga> S com sg kom @ADVL
"<,>"
        "," <,> Z Com <clb>
"<et>"
        "et" <et+0> J sub @J
"<seda>"
        "see" <see+da> P dem sg part @NN>
"<nägu>"
        "nägu" <nägu+0> S com sg part @OBJ
"<oli>"
        "ole" <ole+i> V aux indic impf ps3 sg ps af FinV Intr @FCV @C-OBJ
"<ta>"
        "tema" <tema+0> P pers ps3 sg nom @SUBJ
"<näinud>"
        "näge" <näge+nud> V main partic past ps #Part-P InfP @IMV

"<ennegi>"
        "ennegi" <ennegi+0> D @ADVL
```

Sofie puudepank

- Nordic Treebank Network project
- Jostein Gaarderi novellil "Sopfe maailm" 1. peatükk.
- Rootsi, norra, saksa, taani, islandi, fääri, eesti keeled
  - taani keel: Discontinuous Grammar dependency treebank ja VISL-style phrase structure treebank
  - Swedish: dependency treebank
  - German: NEGRA-style treebank
  - Norwegian: phrase structure treebank
  - Estonian: Penn-style phrase structure treebank.
- Ühine ja paralleelistatud estus TIGER XML kujul
- Eesti osa loodud käsitsi.

Eesti Sofie


S
S
S
"T"
SBJ
TMP
PP
VP
VP
Sofie pani värava enda järel kinni ja tegi kohe ümbriku lahti $.
S_prop V_main S_com P_pos K_post D J_crd V_main D S_com D Z_Fst

Arborest

- Heli Uibo ja Eckhard Bick
- http://corp.hum.sdu.dk/arborest.html
- EstCG korpus konverteeriti VISLPSG reeglite abil funktsioonimärgenditega fraasistruktuuripuudeks
- Märgendus

Arborest-2

- Töö käik
- Lihtlaused
  - ESTCG-formaat teisendada VISLpsg formaati (Tehtud!)
  - rakendada VISLpsg reegleid
  - parandada käsitsi
- Liitlaused
  - ESTCG teisendada ESTCG2
  - ESTCG2 grammatika
  - ESTCG2 teisendada VISLpsg
  - parandada käsitsi

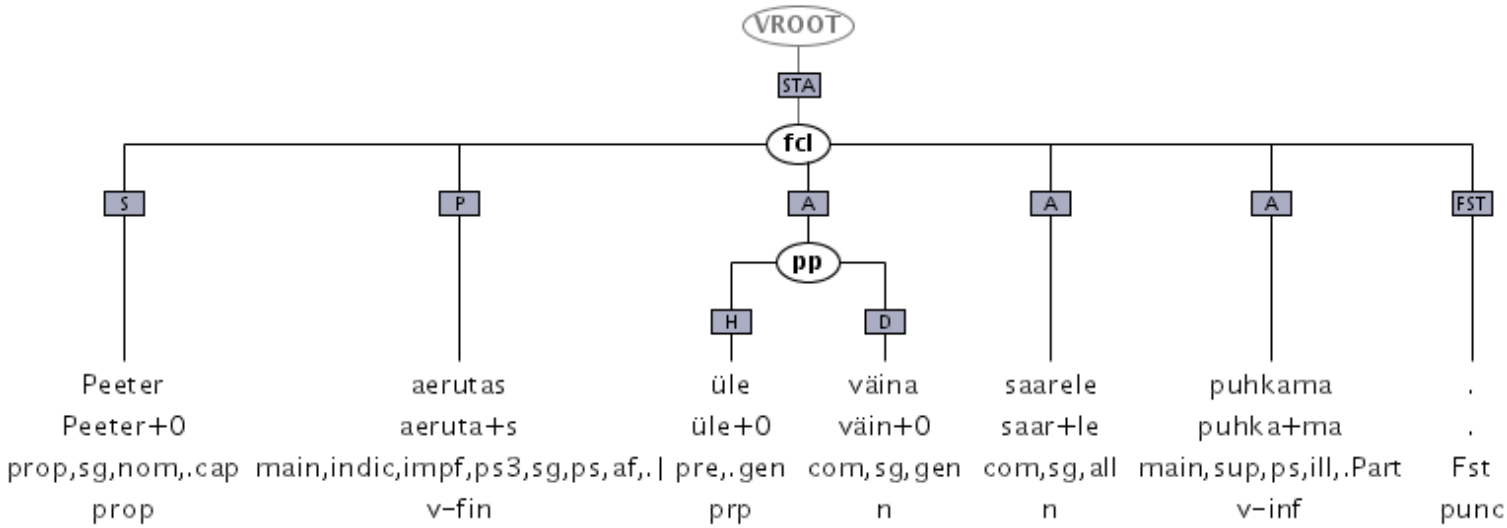
Visualiseerimise ja parandamise vahendid

- Annotate
- Wordfreak
- VISL Tree Editor
- TigerSearch

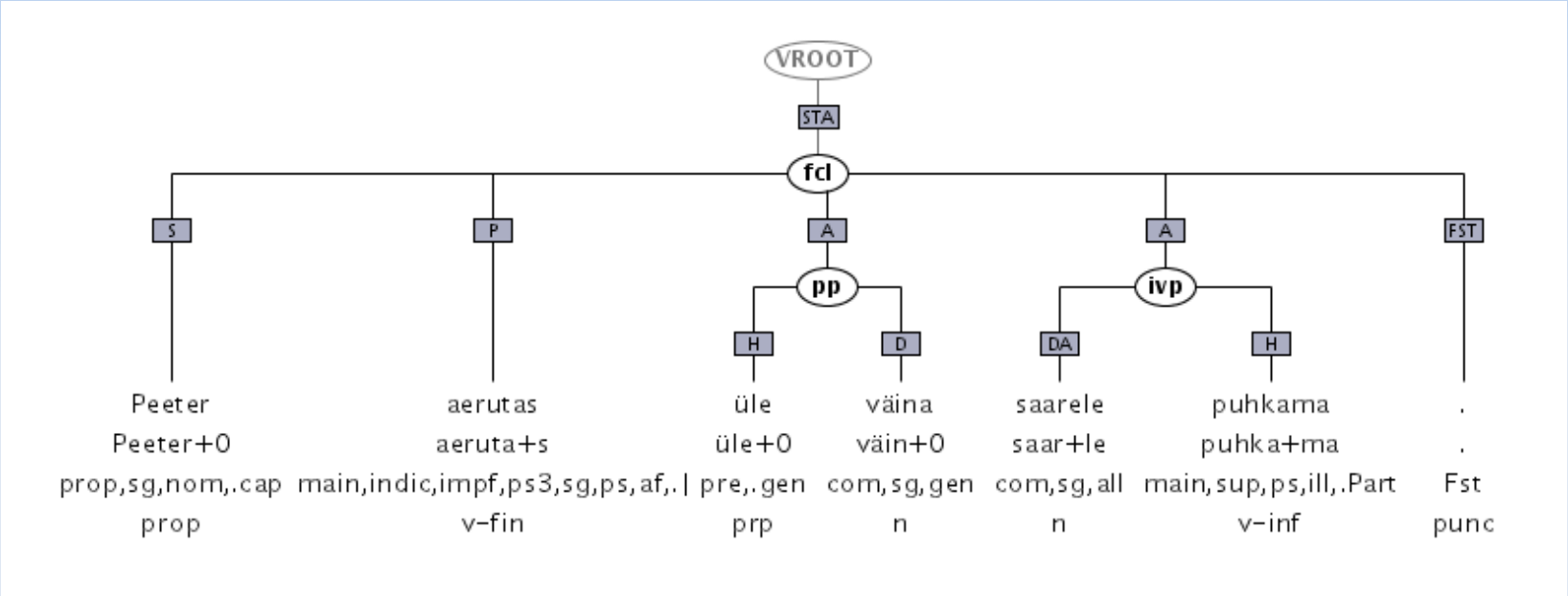
Teisendamine TIGER XMLi

- Olemas skript vislpsg2tigerxml (esialgne versioon E. Bickilt), pole testinud keerulisemaid konstruktsioone.
- Kontrollida märgendust, et see ühilduks kõigega.
- Mida teha metsaga?

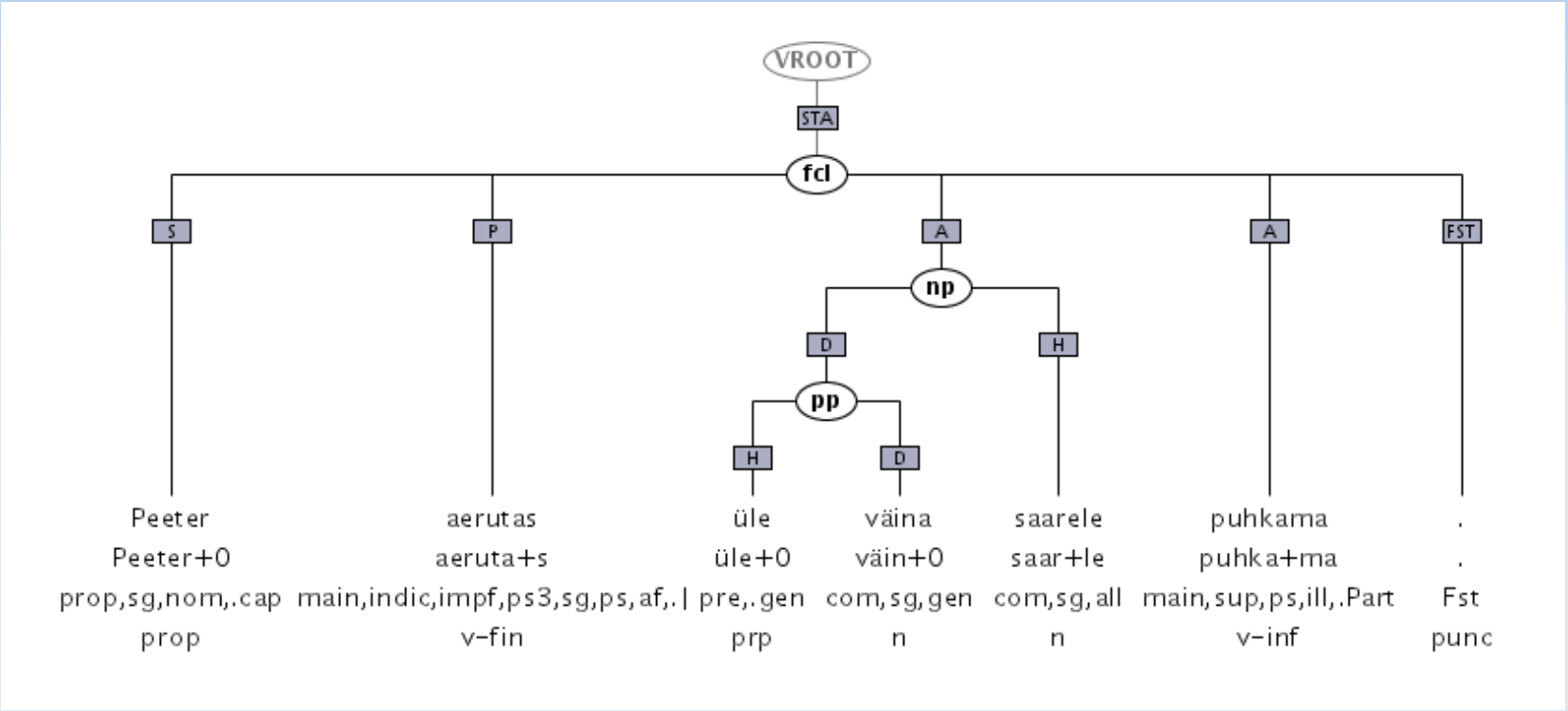
Puu nr 1


VROOT
STA
fcl
S
P
A
A
A
FST
pp
H
D
Peeter
Peeter+0
prop,sg,nom,.cap
prop
aerutas
aeruta+s
main,indic,impf,ps3,sg,ps,af,.|
v-fin
üle
üle+0
pre,.gen
prp
väina
väin+0
com,sg,gen
n
saarele
saar+le
com,sg,all
n
puhkama
puhka+ma
main,sup,ps,ill,.Part
v-inf
.
.
Fst
punc

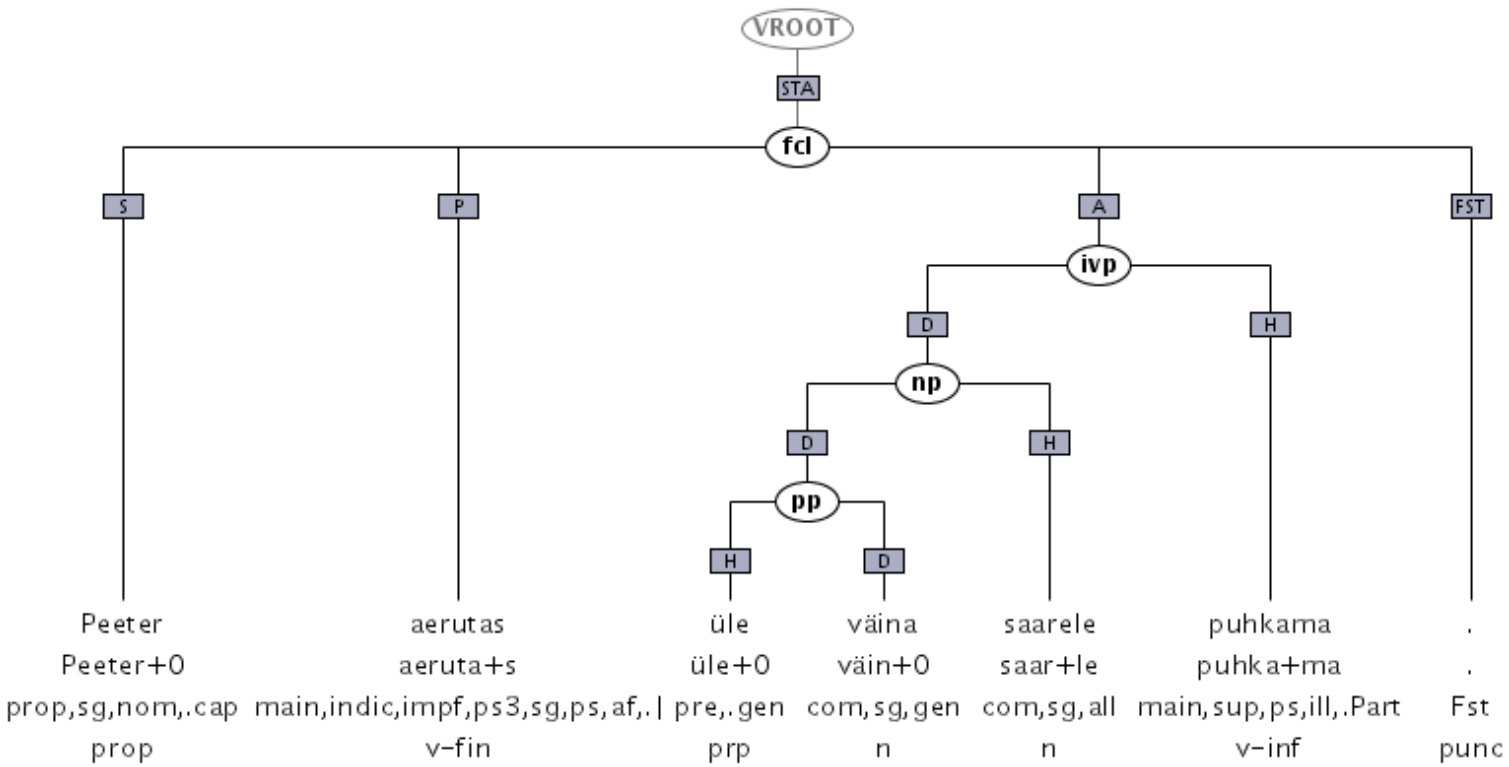
Puu nr 2


VROOT
STA
fcl
S
P
A
A
FST
pp
ivp
H
D
DA
H
Peeter
Peeter+0
prop,sg,nom,.cap
prop
aerutas
aeruta+s
main,indic,impf,ps3,sg,ps,af,.|
v-fin
üle
üle+0
pre,.gen
prp
väina
väin+0
com,sg,gen
n
saarele
saar+le
com,sg,all
n
puhkama
puhka+ma
main,sup,ps,ill,.Part
v-inf
.
.
Fst
punc

Puu nr 3


VROOT
STA
fcl
S
P
A
A
FST
np
D
H
pp
H
D
Peeter
Peeter+0
prop,sg,nom,.cap
prop
aerutas
aeruta+s
main,indic,impf,ps3,sg,ps,af,.|
v-fin
üle
üle+0
pre,.gen
prp
väina
väin+0
com,sg,gen
n
saarele
saar+le
com,sg,all
n
puhkama
puhka+ma
main,sup,ps,ill,.Part
v-inf
.
.
Fst
punc

Puu nr 4


VROOT
STA
fcl
S
P
A
FST
ivp
D
H
np
D
H
pp
H
D
Peeter
Peeter+0
prop,sg,nom,.cap
prop
aerutas
aeruta+s
main,indic,impf,ps3,sg,ps,af,.|
v-fin
üle
üle+0
pre,.gen
prp
väina
väin+0
com,sg,gen
n
saarele
saar+le
com,sg,all
n
puhkama
puhka+ma
main,sup,ps,ill,.Part
v-inf
.
.
Fst
punc