

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

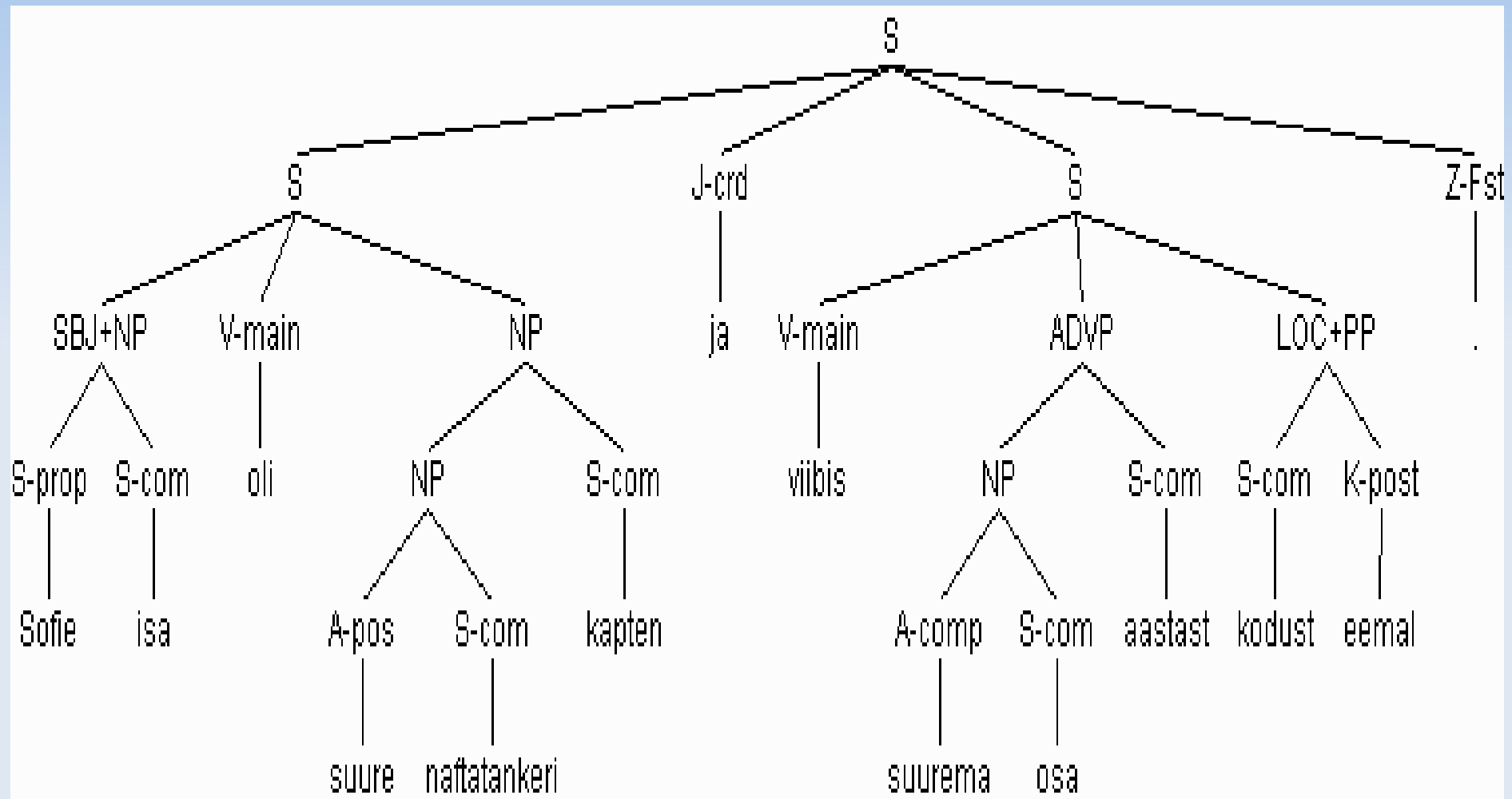
- Tuntuim esitusviis on suluesitus:
 - Lancaster Parsed Corpus

[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)

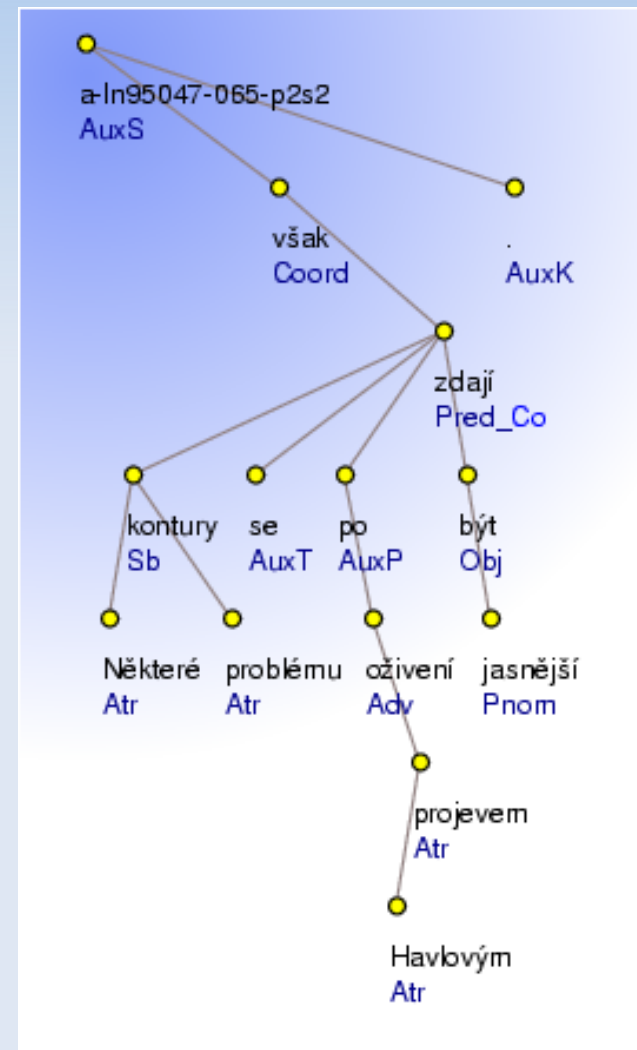
((S
 (NP Martin Marietta Corp.)
 was
 (VP given
 (NP a \$ 29.9 million Air Force contract
 (PP for
 (NP low-altitude navigation and targeting equipment))))))
.)

Puuesitus



Sõltuvuspauude pangad

- Suurim Praha puudepank – 90000 lauset (2005?)



TIGER Treebank

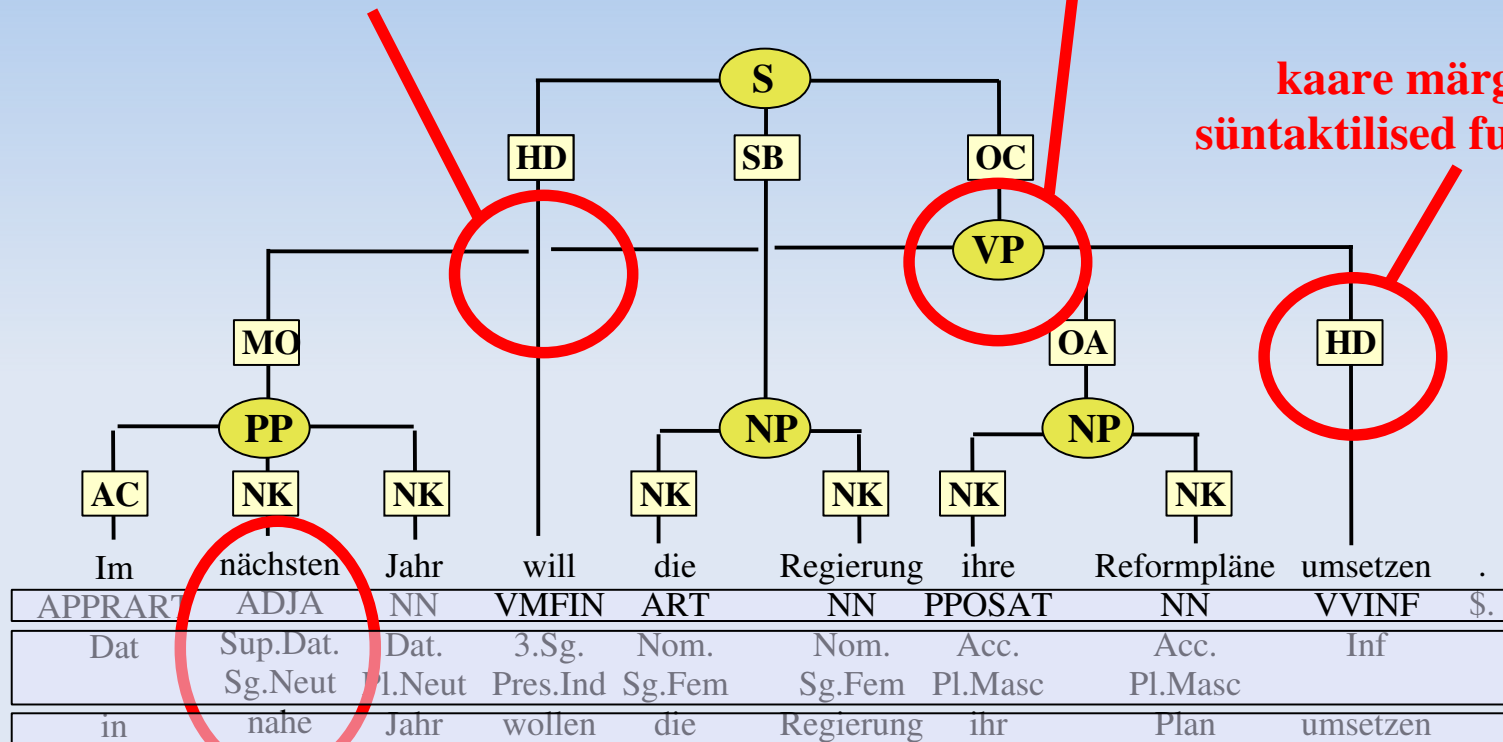
- TIGER saksa keele puudepank – 50000 lauset, koos NEGRA korpuse 20000 lausega võrreldav Penn Treebankiga (1.5 milj sõna)
- **TIGER**: Linguis**T**ic **I**nterpretation of a **GER**man Corpus

TiGer: märgendus

ristuvad kaared

tipu märgendid:
fraasistruktuur

kaare märgendid:
süntaktilised funktsioonid



sõna tasandil märgendus:
sõnaliik, morfoloogiline info
lemma

VISL treebanks

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

Arborest

Syntax Learning

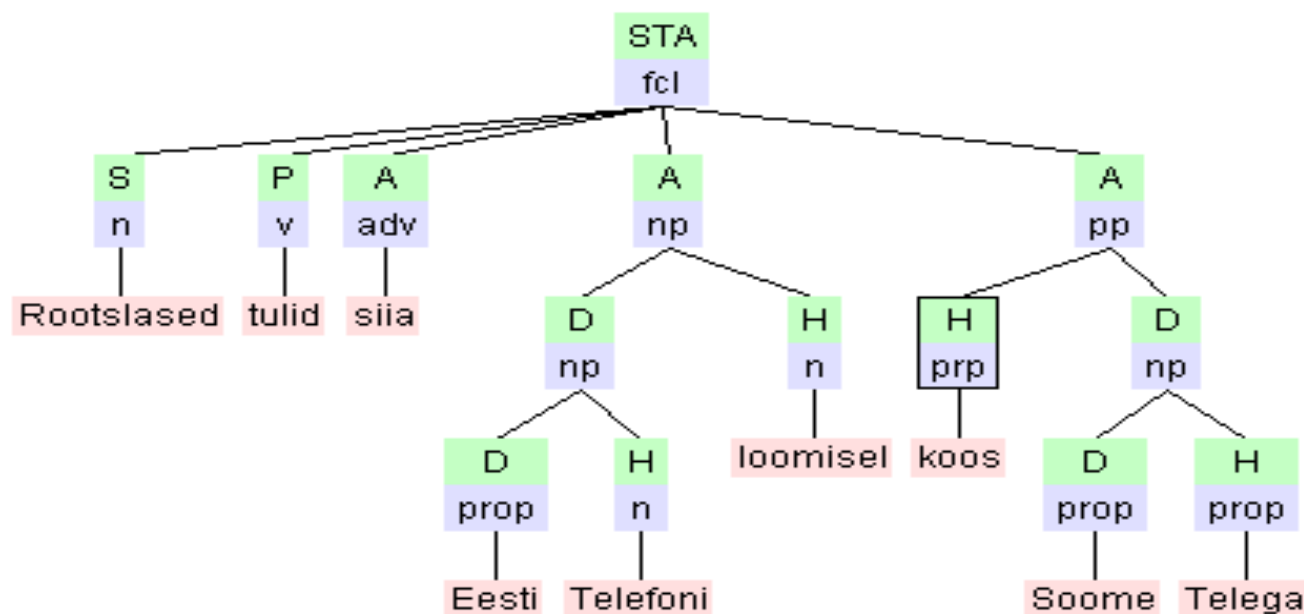
Language Settings Tools Help

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)



Arboreset puu formaat

SOURCE id='est-125'

ID='est-125' Rootslosed tulid siia Eesti Telefoni loomisel koos
Soome Telega.

A1/2

STA:fcl

S:n("rootslane+d" com pl nom .cap) Rootslosed

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>"

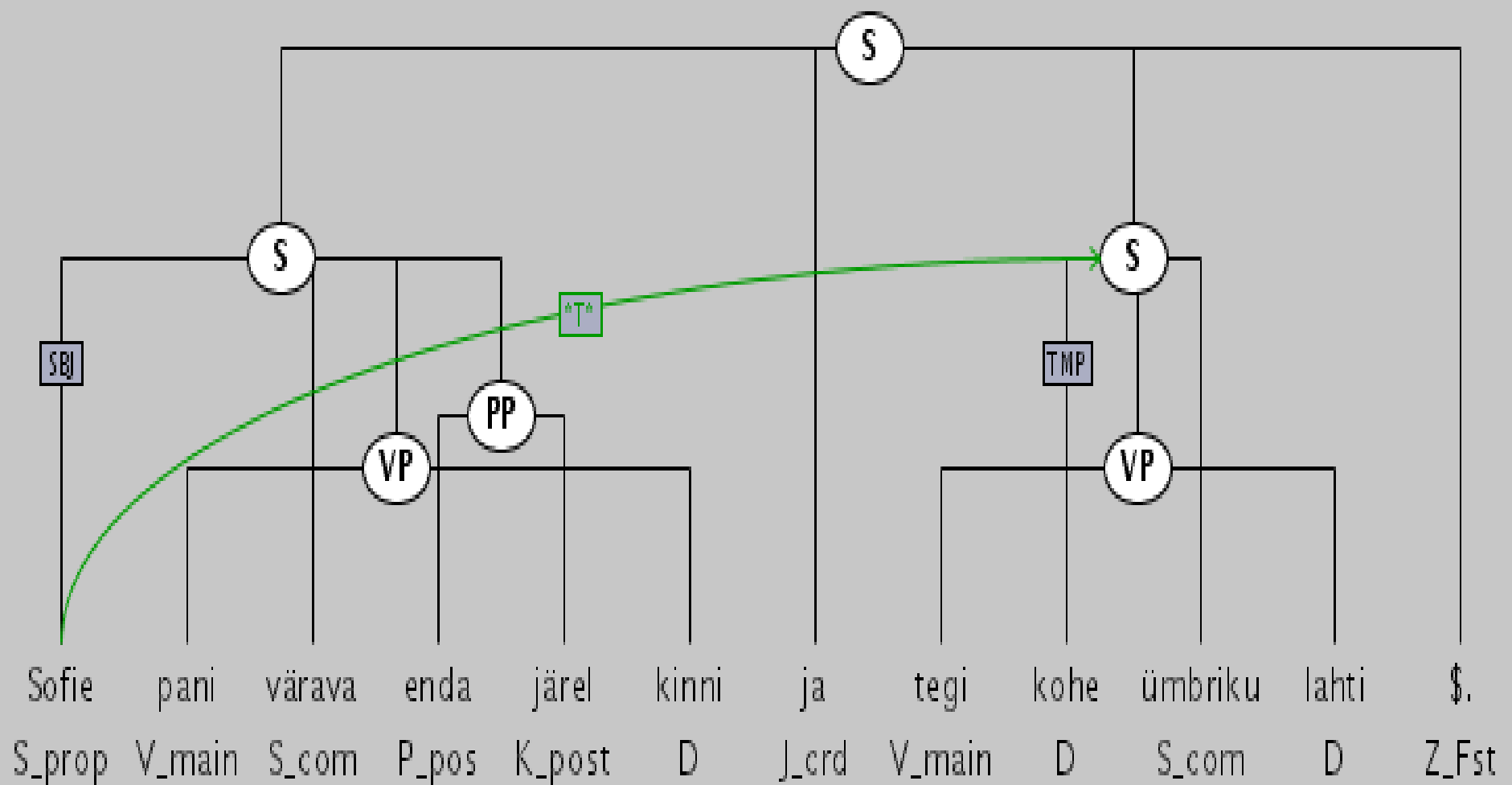
"näge" <näge+nud> V main partic past ps #Part-P InfP @IMV

"<ennegi>"

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



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

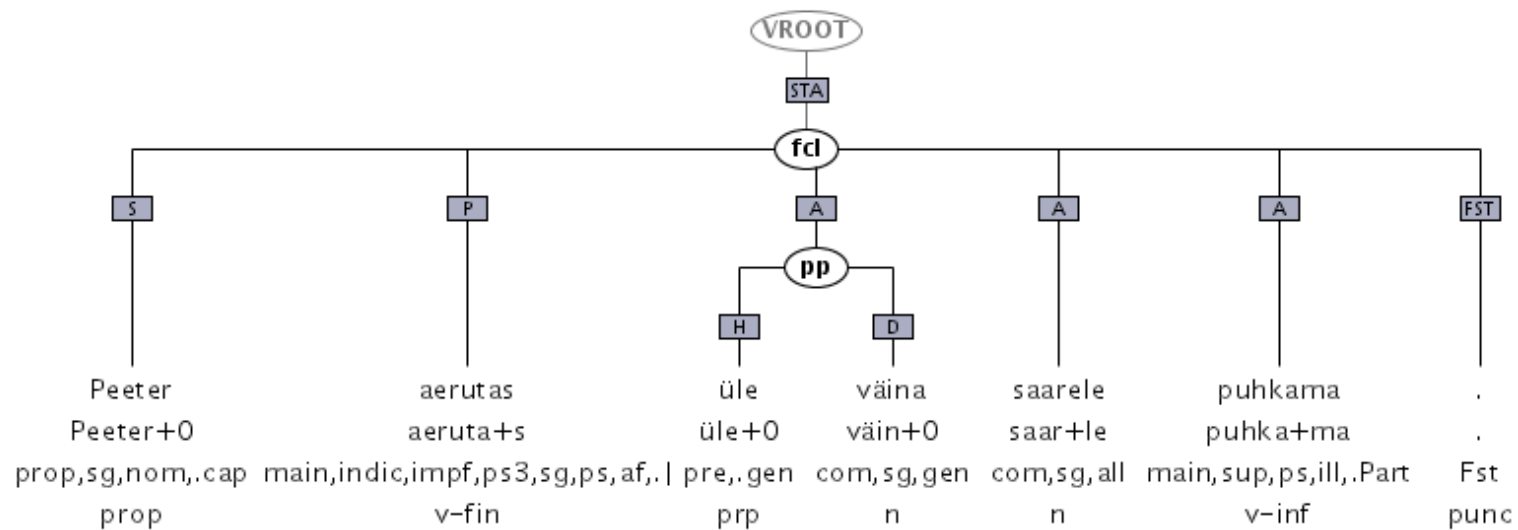
Visualiseerimise ja parandamise vahendid

- Annotate
- Wordfreak
- [VISL Tree Editor](#)
- TigerSearch

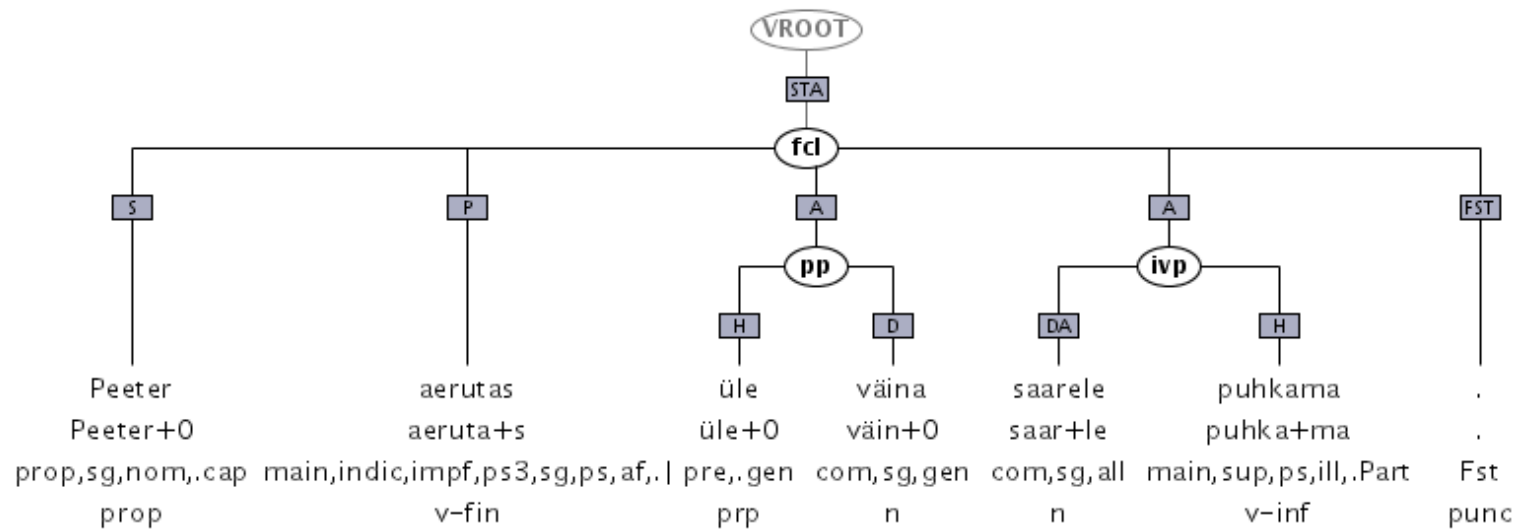
Teisendamine TIGER XMLi

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

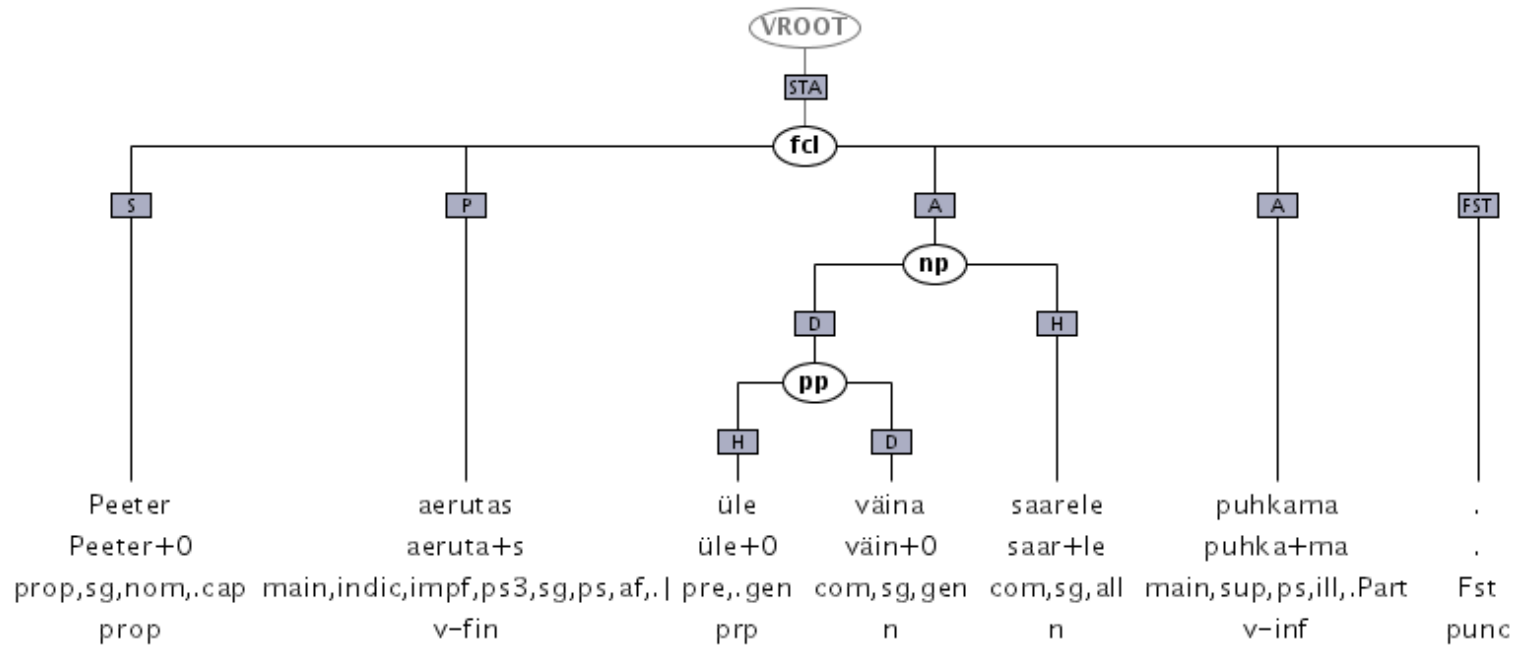
Puu nr 1



Puu nr 2



Puu nr 3



Puu nr 4

