

Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid

Tiit Roosmaa, Kaili Müürisep, Helen Nigol,
Heli Uibo, Kaarel Kaljurand

Ettekande ülevaade

- Projekti eesmärgid
- Milleks süntaksianalüsaator?
- Sügav analüüs
- Suulise keele süntaktiline analüüs
- Sisukokkuvõtja
- Grammatikakorrektor

Projekti eesmärgid

- luuakse järgmiste keeletarkvarasüsteemide prototüübid:
 - grammatikakorrektor
 - süntaksianalüüsil põhinev automaatsete sisukokkuvõtete tegija
 - süntaksianalüüsil põhinev infootsisüsteem
- Nimetatud keeletarkvara prototüüpide loomiseks ja testimiseks on vaja pind- ja süvasüntaktiliselt märgendatud treening- ja testkorpusi.
- Eesti keele süntaksipuude panga märgendus peaks olema ühilduv või teisendusrelatsioonis Põhjamaade paralleelpuudepanga märgendamiseks valitava formalismiga.

Süntaksianalüsaator - milleks?

- Praktilised rakendused:
 - grammatikakorrektor
 - masintõlge, tõlkemälu
 - teised “sisu mõistvad” rakendused:
 - sisukokkuvõtja,
 - dialoogsüsteemid
 - õpiprogrammid
- Teoreetilised rakendused:
 - süntaktiliselt märgendatud korpused ja puudepangad
 - sisend semantilisele analüüsile

Korpus ja puudepank

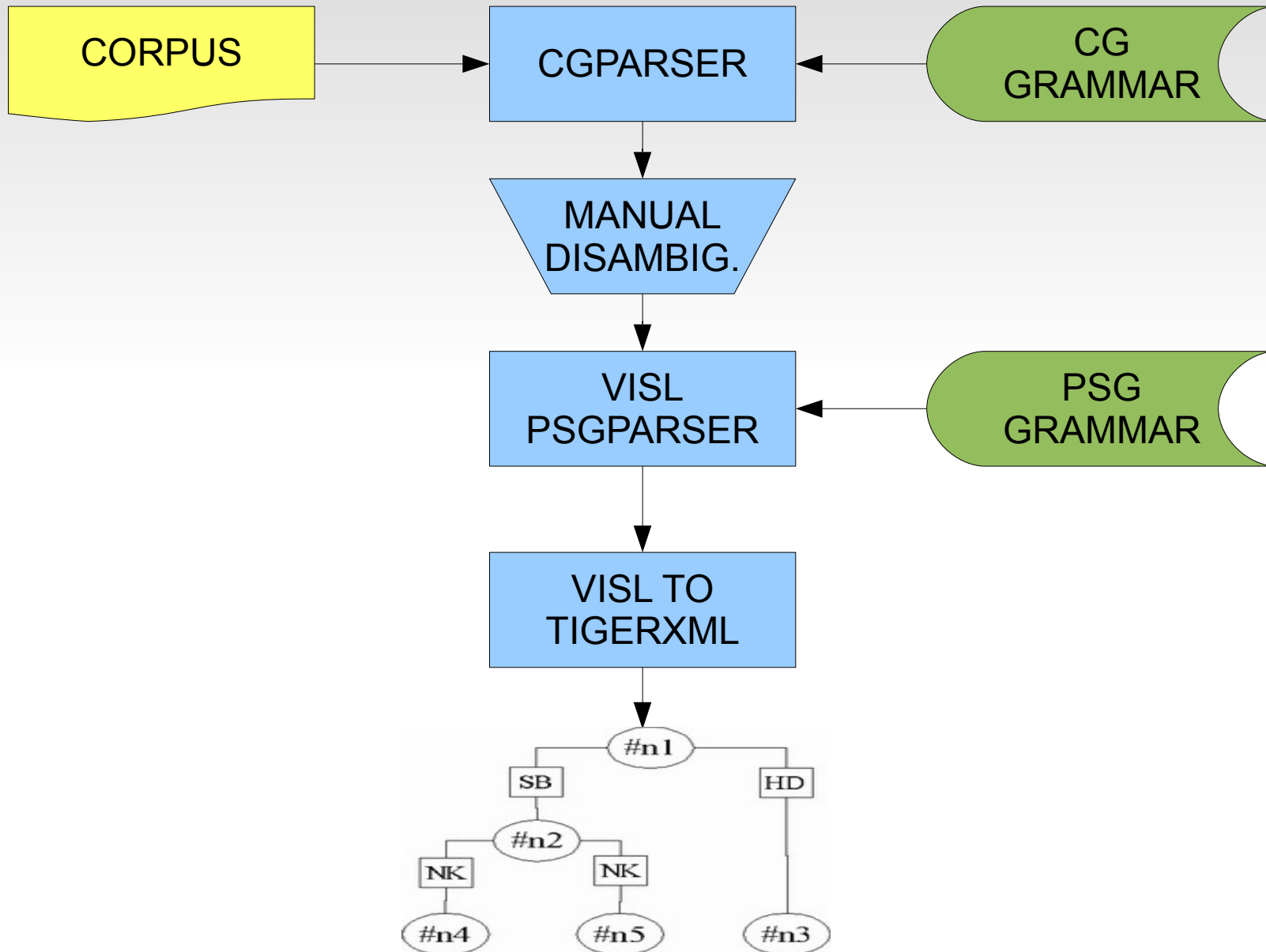
- Loodud ligikaudu poole miljoni sõnaline süntaktiliselt märgendatud tekstikorpus.

täheldades tähelda+des // _V_ main ger af #Part // **CLB @ADVL
Uruguay Uruguay+0 // _S_ prop sg gen #cap // @NN>
vooru voor+0 // _S_ com sg gen // @NN>
läbirääkimistest läbi_rääkimine+test // _S_ com pl el // @ADVL
tulenevaid tulene=v+id // _A_ pos pl part #v partic // @VN>
kohustusi kohustus+i // _S_ com pl part // @OBJ
\$, , // _Z_ Com //
mis mis+0 // _P_ inter rel pl nom // **CLB @SUBJ
puudutavad puuduta+vad // _V_ main indic pres ps3 pl ps af // @+FMV
füüsiliste füüsiline+te // _A_ pos pl gen #line // @AN>
isikute isik+te // _S_ com pl gen // @NN>
liikumist liiku=mine+t // _S_ com sg part #mine // @OBJ
teenuste teenus+te // _S_ com pl gen // @NN>
osutamiseks osuta=mine+ks // _S_ com sg tr #mine // @<NN

Sügavam analüüs

- Leida lause puukujuline struktuur
- 370 liikumisverbiga lihtlauset H. Rätsepa näitelauseste korpusest, mis sobivad semantilise analüüsi tööruhmale

Analüüsi skeem



Lause pindmine analüüs

\$<s>

Peeter

Peeter+0 // _S_ prop sg nom #cap // @SUBJ

hiilis

hiili+s // _V_ main indic impf ps3 sg ps af #FinV // @+FMV

linnaääri

linna_äär+i // _S_ com pl part // @P>

mööda

mööda+0 // _K_ post #part // @ADVL

koosolekult

koos_olek+lt // _S_ com sg abl // @ADVL

koju

kodu+0 // _S_ com sg adit // @ADVL @NN> @<NN

püssi

püss+0 // _S_ com sg gen // @P>

järele

järele+0 // _K_ post #gen // @ADVL

.

. // _Z_ Fst //

\$</s>

!<hiilima#67.3.>

VISL PSGparser

ADJP:ap = ADVL[->D]:adv AN>[->H];

ATRNP:np = {AN>,ADJP}[->D] NN>[->H];

S:np = {AN>,ADJP,NN>,ATRNP}[->D] S[->H];

S:np = S[->H] <NN[->D];

PREP:np = {AN>,ADJP,NN>,ATRNP}[->D] P<[->H];

A:pp = A[->H]:prp <P[->D];

A:pp = A[->H]:prp PREP[->D]:np;STA:fcl = S P A O A A FST;

STA:fcl = S P O A FST;

STA:fcl = S P A O A FST;

STA:fcl = S P O A A FST;

STA:fcl = S P A O A FST;

Visli formaat

STA:fcl

S:prop('Peeter+0',prop,sg,nom,.cap) Peeter

P:v-fin('hiili+s',main,indic,impf,ps3,sg,ps,af,.FinV) hiilis

A:pp

=D:n('linna-ää;r+i',com,pl,part) linnaääri

=H:pst('mööda+0',post,.part) mööda

A:n('koos-olek+lt',com,sg,abl) koosolekult

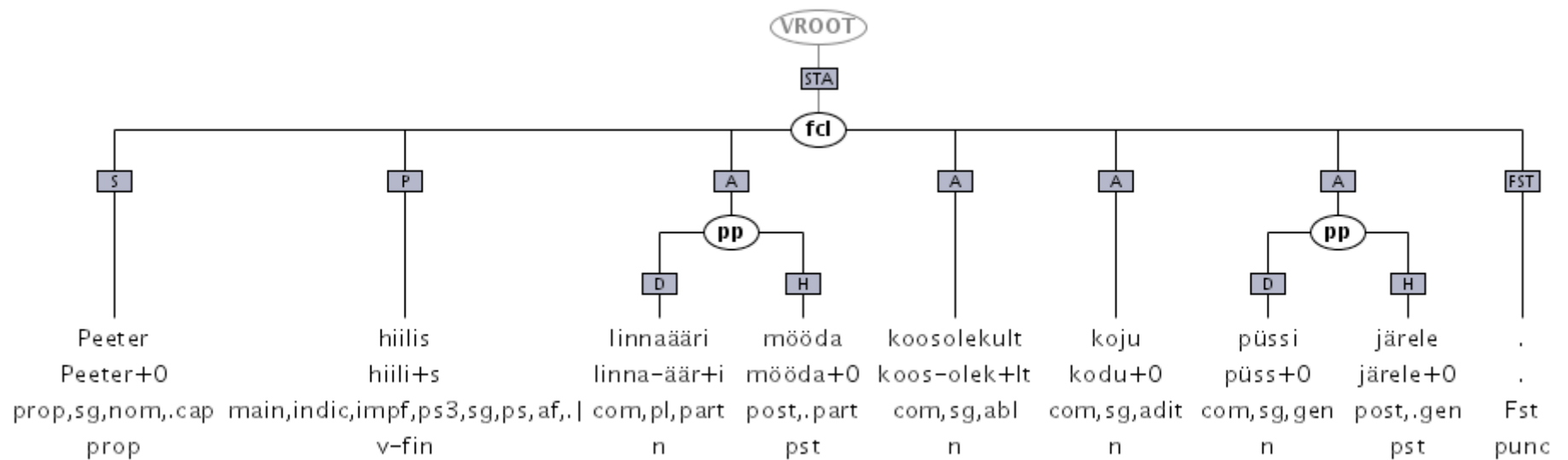
A:n('kodu+0',com,sg,adit) koju

A:pp

=D:n('püss+0',com,sg,gen) püssi

=H:pst('järele+0',post,.gen) järele

TigerXML



Suulise keele analüüs

H: [ee:] oskate ´oelda kas ´homme oleks võimalik pool´teljed Renool ´ära vahetada.

V2: m ´homme ´kindlasti ei ole võimalik seda ´as[ja]

H: [ei]=´ole=jah (1.0)

V2: jah (.) ja mis ´auto teil ´ültse [on {---} ei=no]

H: [RE´NOO RE´NOO. Renoo üks´teist.]

V2: ja=ja=ja mis prob´leem teil on. üks[´teist]=ä

H: [ta]

H: jah tal on: mõlemad pool ´teljed nagu vaja ´ära vahetada.

Uued märgendatud korpused

- 8400 sõna treeningkorpus
- 6700 sõna testkorpus
- 13000 sõna mitteladususte korpus
 - Parandused
 - Kordused
 - Valestardid

Valestardid

K #####

\$<s>

muna muna+0 //_S_ com sg nom // **CLB @SUBJ

noh noh+0 //_B_ // @B

see see+0 //_P_ dem sg nom // @<NN

siia siia+0 //_D_ // @ADV

asemele asemele+0 //_D_ // @ADV

tuleks tule+ks //_V_ main cond pres ps3 sg ps af #FinV #Intr //@+FMV

leida leid+a //_V_ main inf #NGP-P // @OBJ

midagi miski+dagi //_P_ indef sg part // @OBJ

muud muu+d //_P_ indef sg part // @<NN

ma mina+0 //_P_ pers ps1 sg nom // **CLB-C @SUBJ

soovitaks soovita+ks //_V_ main cond pres ps1 sg ps af // @+FMV

hapukoort hapu_koor+t //_S_ com sg part // @OBJ

\$. . //_Z_ Fst //

Tulemused

	Kirjalik	Enne	Nüüd
Sõnu		2543	6717
Saagis	98.5	97.3	97.7
Täpsus	87.5	89.2	90.4
Ühesus	89.5	91.5	93.0

Mis on sisukokkuvõtja

- Programm, mis teeb olemasolevast tekstist lühema versiooni, esitades kasutajale ainult vajalikku infot.
- Sisukokkuvõtte põhieesmärk on esitada teksti peamised ideed väiksemas mahus.

EstSumist

- Genereerib väljavõtte
- Väga pindmine meetod ja lihtne algoritm
- Perli-programm
- Eelkõige ajaleheartiklid (uudised)

Arhitektuur

- EstSum koosneb kolmest moodulist:
 - HTML-konverter,
 - lausestaja,
 - väljavõtete tegija.

Ekstraktor

- $W(s) = \alpha P(s) + \beta F(s) + \gamma K(s)$
- $P(s)$ – positsiooniskoor
 - Teksti 1. lause
 - Lõigu esimene lause
 - Lõigu teine ja kolmas lause
- $F(s)$ – formaadiskoor
- $K(s)$ – Võtmesõnade skoor

Ekstraktor - näide

<div0 type='unknown'><head>**Vabadussamba kavand sai
sihvakust juurde**</head>

<p>

1. p=16.129032 f=5.154639 w=2.679067 s=9.049282 <s>**Eile Tallinna
kunstihoones Vabadussõja võidusamba täiustatud kavandit
esitlenud autorid avaldasid veendumust, et viimaks on leitud
kunstiliselt ja linnaehituslikult parim variant, mida pole põhjust
enam muuta.**</s>

</p>

<p>

8. p=6.451613 f=5.154639 w=2.548987 s=5.152298 <s>**Võrreldes
algse kavandiga on autorid põhjalikult ümber kujundanud ka
Vabaduse väljaku servale rajatava monumendi
tseremooniaväljaku, mida hakkavad ilmestama tribüünina või**

Tulemused

- Kuidas hinnata sisukokkuvõtte headust? Mis teeb ühe sisukokkuvõtte heaks ja teise halvaks?
- Kahe inimese poolt koostatud väljavõtetes kattub ainult 70% lausetest
- EstSumi poolt valitud laused kattusid 60% ulatuses inimese poolt valitud lausetega. Parimal juhul oli samu lauseid 85% ja halvimal juhul ei kattunud ükski

Positiivne näide

<div0 type='unknown'><head>**Tänavune beebisaak ületab 15 000 piiri**</head>

<p>

1. $p=19.083969$ $f=4.464286$ $w=2.659838$ $s=9.951270$ <s>**Eelmisel aastal jäi ilus ümmargune number - 15 000 vastündinut - napilt saavutamata.**</s>

2. $p=5.343511$ $f=4.464286$ $w=5.107198$ $s=4.944558$ <s>**Puudu jäi vaid 123 last.**</s>

</p>

<p>

23. $p=7.633588$ $f=4.464286$ $w=1.915900$ $s=5.222330$ <s>**Lisaks üksikute laste suuremale sündimusele näeb aasta aastalt ilmavalgust üha enam mitmikke, milles on oma osa ka kunstlikul viljastamisel.**</s>

Negatiivne näide

AK uus graafiline lahendus: diletantism riigitelevisioonis

1. mail eetrisse paisatud «Aktuaalse kaamera» uue kujunduse kohta võib pressiteatest lugeda: «Koostöös reklaamiagentuuriga Kontuur töötati välja ka uus AK logo.»

ETV pearepissöör René Vilbre ütleb: »Lisaks uuele kujundusele muudame ka AK logo ehk graafilist ühendit AK. Iga logo peaks olema arusaadav, lihtne ja selgelt märgikeskne.» Märksõnana nimetatakse veel «uudiste kiirteed».

Midagi on sassi läinud.

Esiteks.

Teiseks. Kontuur on üks Eesti vanemaid ja kogenumaid reklaamiagentuure, loendamatute õnnestunud logode looja.

Kolmandaks.

Neljandaks. Lihtsuse ja minimalismi asemel võiks pigem rääkida hõredusest ja abitusest.

Tehtust õhkub diletantismi ja põhjendamatust. Kasutatud kirjatüüp Bank Gothic

Õigekirjakorrektor või grammatikakorrektor

- Õigekirjakorrektor kontrollib, kas sõna leidub leksikonis kontekstile tähelepanu pööramata.
- Avastab ilmsed vead: krammatika, kogematta jne
- Grammatikakorrektor vaatab kogu lauset.
- Leiab (ideaaljuhul) sõnad, mille vorm ei sobi lausesesse (muutis elu praemaks),
- Ühildumis- ja rektsioonivead: ta võis tegema
- Komavead

Komavigade korpus

- Kogume komavigasid
 - tavalised korpuse tekstid, kust on kirjavahemärgid eemaldatud
 - tekstid, mis on hooletult toimetatud
 - spontaanselt kirjutatud tekstid foorumitest ja kommentaariumitest

Idee

- Kasutades kitsenduste grammatika analüsaatori mootorit märgendada kahtlased sõnad märgendiga "korrektne" või "vigane"

"<Soovitan>" "soovita" <soovita+n> V main indic pres ps1 sg ps af .NGP-
P .InfP **CLB +FMV

"<kõikidel>" "kõik" <kõik+del> P det pl ad ADVL

"<kes>" "kes" <kes+0> P inter rel pl nom **CLB-C SUBJ OBJ @ERR

"<sellist>" "selline" <selline+t> P dem sg part NN>

"<teed>" "tee" <tee+d> S com sg part OBJ

"<näinudki>" "näge" <näge+nudki> V main partic past ps .Part-P .InfP
-FMV

"<pole>" "ole" <ole+0> V aux indic pres ps neg .FinV .Intr +FCV

"<, >" ", " <, > Z Com

Plaanid

- Liitlausete sügavam analüüs
- Kõnekonaruste tuvastamine
- Suulise keele puudepanga formaat
- Sisukokkuvõtja uus versioon
- Grammatikakorrektori komavigade tuvastaja prototüüp ja hinnang efektiivsusele