Automaatne süntaktiline analüüs

Kaili Müürisep
Tartu Ülikool
Arvutiteaduse instituut

Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid 2006-2008

Eesmärgid:

- Luua keeletarkvara prototüübid grammatikakorrektorile ja automaatsele sisukokkuvõtete tegijale
- Luua selleks vajalikud keeleressursid

Pindsüntaktiliselt märgendatud korpus

Loodud ligikaudu poole miljoni sõnaline süntaktiliselt märgendatud tekstikorpus:

- Ilukirjandus
- Tõlkekirjandus
- Ajakirjandus
- Seadus
- Verbirektsioonidega lihtlaused
- Suuline keel (argivestlus, infodialoog)
- Murre

Pindsüntaktiliselt märgendatud korpus

$<s>
See
    see+0 //_P_ dem sg nom #cap // **CLB @SUBJ
oli
    ole+i //_V_ main indic impf ps3 sg ps af #Intr // @+FMV
osa
    osa+0 //_S_ com sg nom // @PRD
vihkamise
    vihka=mine+0 //_S_ com sg gen #mine // @NN>
nädala
    nädal+0 //_S_ com sg gen // @ADVL
eelsest
    eelne+st //_A_ pos sg el // @AN>
kokkuhoiukampaaniast
    kokku_hoiu_kampaania+st //_S_ com sg el // @<NN
$.
    $. //_Z_ Fst //
$</s>

Suulise keele süntaktilisest märgendamisest

T: ja noh tähendab `mina sellest `aru ei saa sest=noh mina `lõikangi `vasaku `käega [`kääride]ga, mul=ei ma=ei tuld `selle `pealegi=t võib mingi `erinevus olla. (.) [ma]=aint `mõtsin need on `õutselt `mugavad. [(0.8) et=nad]

L: [mh] [((naerab))]

T: `niigi sobivad `kätte. [(0.5) `pöid]la ja [kõik on] `tehtud.=

Kõnekonarused

K ####

$<s>

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

noh noh+0 //_B_ // @B I

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

siia siia+0 //_D_ // @ADVL

asemele asemele+0 //_D_ // @ADVL

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 //

$</s>

Murdetekst

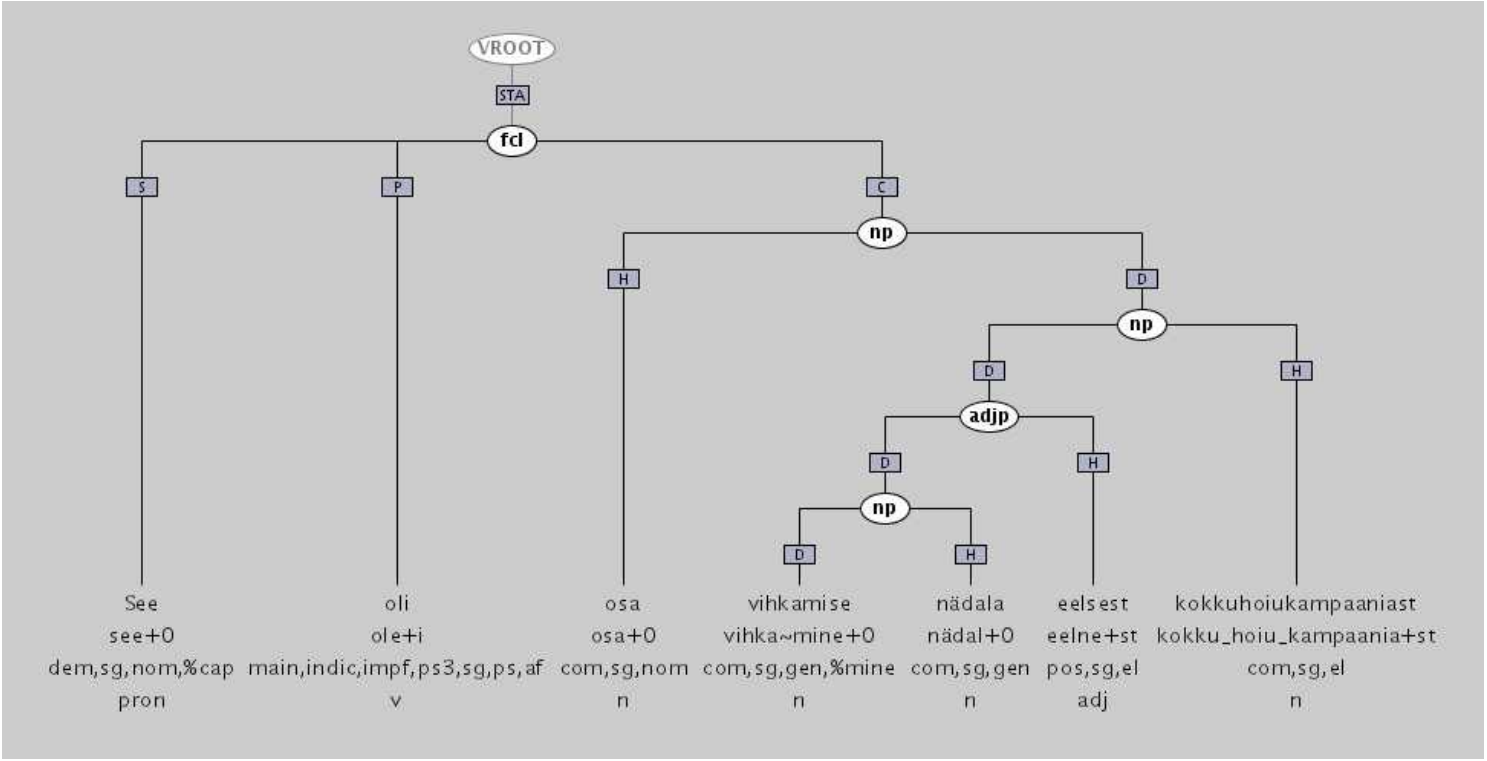
```
$.$.$.
    $. //_Z_ Fst //
siss [ADVL]
    siis+0 //_D_ //    **CLB @ADVL
nakkatti [+FMV]
    nakka+0 //_V_ main imps indic impf #FinV #Intr #All //   @+FMV
mullõ [ADVL]
    mina+0 //_P_ pers ps1 sg all //   @ADVL
õks [B]
    iks+0 //_B_ //   @B
`palkka [OBJ]
    palk+0 //_S_ com sg part //   @OBJ
ka [B]
    ka+0 //_B_ //   @B
`masma [-FMV]
    maks+0 //_V_ main sup ill //   @-FMV @ADVL
```

Puudepank

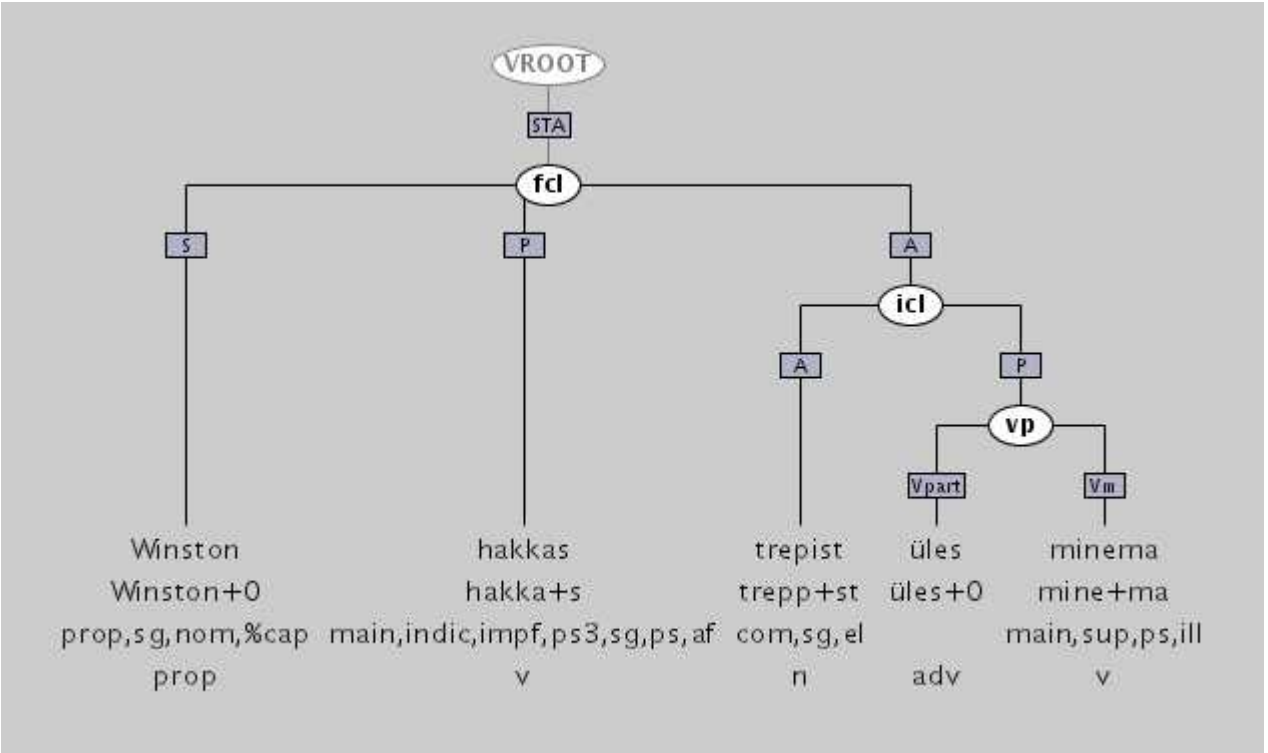
Kogu käsitsi märgendatud puudepank koosneb hetkel:

- 388 liikumisverbiga lihtlauset Rätsepa korpusest.
- 732 liikumisverbiga lauset eesti frameneti testkorpusest
- 175 lauset Arboresti korpusest
- 20 lauset suulise keele korpusest

Puudepank


VROOT
STA
fcl
S
P
C
np
H
D
np
D
H
adjp
D
H
np
D
H
See
see+0
dem,sg,nom,%cap
pron
oli
ole+i
main,indic,impf,ps3,sg,ps,af
v
osa
osa+0
com,sg,nom
n
vihkamise
vihka~mine+0
com,sg,gen,%mine
n
nädala
nädal+0
com,sg,gen
n
eelsest
eelne+st
pos,sg,el
adj
kokkuhoiukampaaniast
kokku_hoiu_kampaania+st
com,sg,el
n

Puudepank


VROOT
STA
fcl
S
P
A
icl
A
P
vp
Vpart
Vm
Winston
Winston+0
prop,sg,nom,%cap
prop
hakkas
hakka+s
main,indic,impf,ps3,sg,ps,af
v
trepist
trepp+st
com,sg,el
n
üles
üles+0
adv
minema
mine+ma
main,sup,ps,ill
v

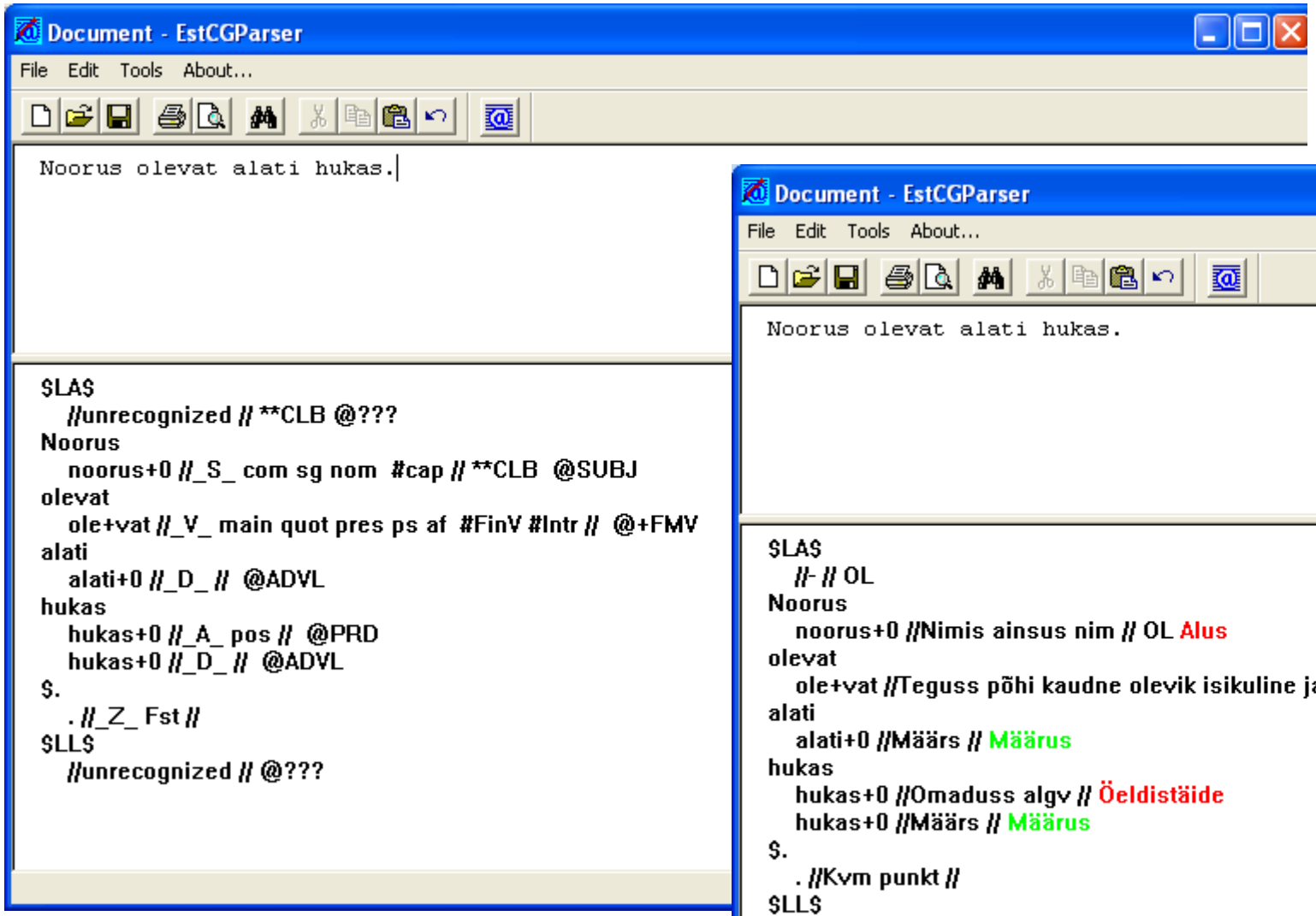
Grammatikad

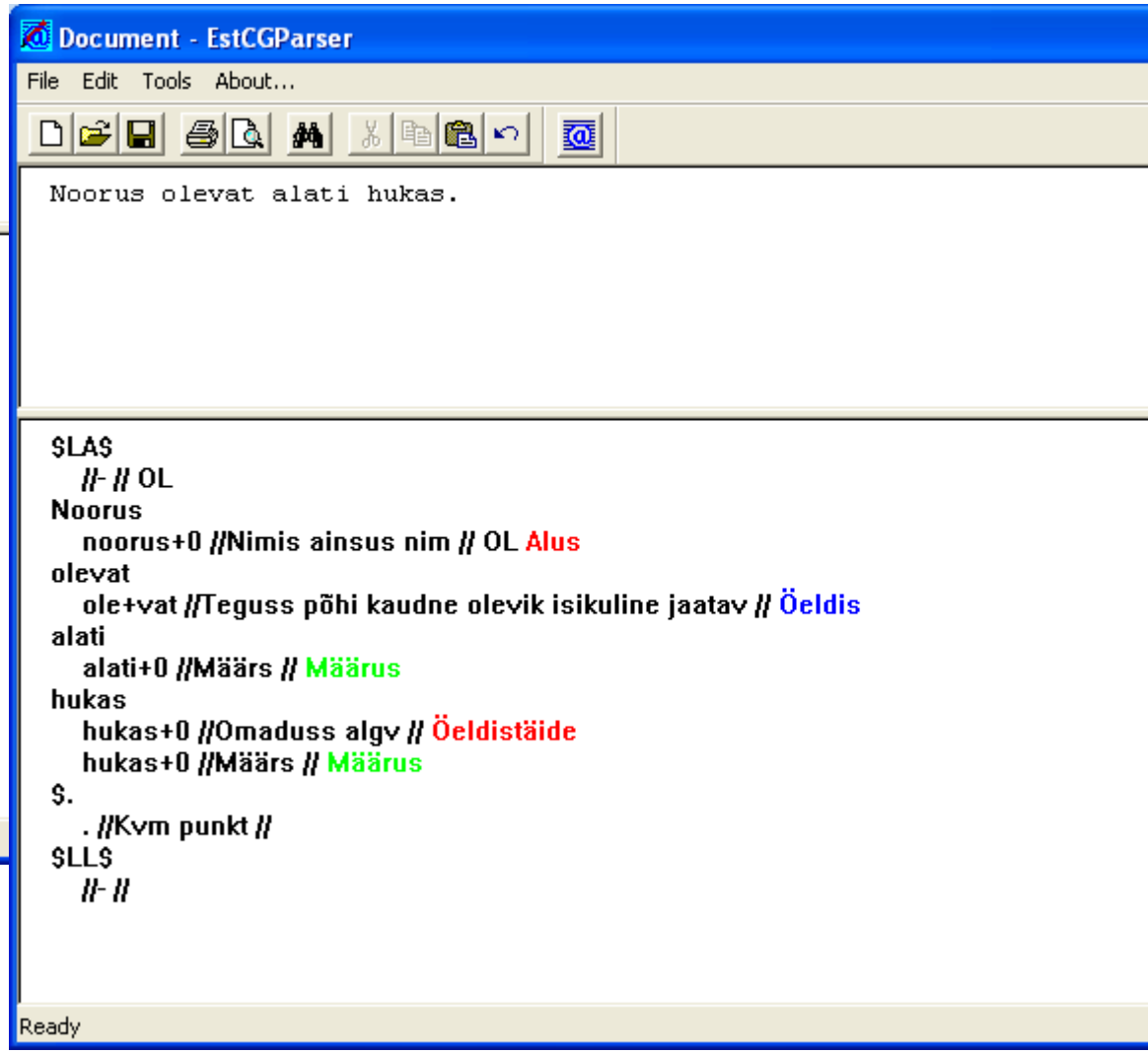
- Teisendati T. Puolakaineni loodud (Puolakainen 2001) morfoloogilise ühestaja reeglid uue VISL parseri formaati (1509 reeglit).
- Teisendati pindmise süntaksianalüsaatori reeglid uue VISL parseri formaati (1130 reeglit).
- Loodi sõltuvusstruktuuri ehitav grammatika (50 reeglit), mille abil on võimalik leida enamiku lausete osalist puustruktuuri.
- Loodi lihtsamate fraasitüüpide analüüsimiseks fraasistruktuurigrammatika, mis siiski ei sobi keerulisemate infiniittarinditega lihtlausete ja liitlausete analüüsiks.

Sõltuvusseosed

```
"<Kuigi>"
    "kuigi" <kuigi+0> J sub <cap>  @J #1->3
    "kuigi" <kuigi+0> D <cap> @ADVL #1->3
"<president>"
    "president" <president+0> S com sg nom  @SUBJ #2->3
"<andis>"
    "and" <and+is> V main indic impf ps3 sg ps af <FinV> <NGP-P> <InfP>
     @FMV #3->11
"<stuudiosse>"
    "stuudio" <stuudio+sse> S com sg ill  @ADVL #4->5
"<saabudes>"
    "saabu" <saabu+des> V main ger <Intr> <Ill>  @ADVL #5->3
"<oma>"
   "oma" <oma+0> P det pos refl sg gen  @NN> #6->7
"<rivaalile>"
   "rivaal" <rivaal+le> S com sg all  @ADVL #7->3
"<kätt>"
    "käsi" <käsi+tt> S com sg part  @OBJ #8->3
```

Windowsi parser





Sisukokkuvõtja

- http://math.ut.ee/~kaili/estsum/2009/estsumframe.cgi

Grammatikakorrektor

- Tugineb morfoloogilise ühestaja ja süntaksianalüsaatori väljundile
- Esimese sammuna lisatakse kaks märgendit, @ERR ja @OK, vastavalt vigase ning korrektse koha märgendamiseks, kõigile sidesõnadele ja verbi pöördelistele vormidele.
-

Näide

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

Grammatikakorrektori reeglid

- Kui sõnale *siis* järgneb vahetult sõna *kui*, siis nõuda nende vahele koma. Nt *Tulen siis, kui tahan.*
- Kui sõnale *siis* eelneb lauses sõna *kui* ning järgneb võimalik öeldis, siis juhul, kui sõnade *kui* ja *siis* vahel leidub võimalik öeldis, nõuda *siis* ette koma. Nt *Kui teha, siis teha hästi.* Vastasel juhul koma mitte panna. Nt J*a kui siis hakkas sadama.*
- Kui järjest esinevad sõnad *kui*, *siis* ja küsisõna, siis nõuda *siis* ette koma. Nt *Ja kui, siis mida*?

Tulemused

- Korrektselt tuvastatud komavigade osakaal kõigist grammatikakorrektori poolt antud komavigade märgenditest ehk grammatikakorrektori täpsus testkorpusel on 93,8%.
- Saagis ehk leitud komavigade suhe korpuses leidunud (ja käsitsi märgendatud) komavigadesse on 94,1%.
- 30 lause hulgas oli 27 komaveaga, neist leiti 24

Projektiga seotud magistritööd

- Aivi Kaljuvee. Määruste ja täiendite eristamine statistiliste meetoditega. Tartu Ülikool, Arvutiteaduse instituut. 2008
- Krista Liin. Reeglipõhine komavigade tuvastaja eestikeelsetele tekstidele. Tartu Ülikool, Arvutiteaduse instituut. 2008
- Pilleriin Mutso. Knowledge-poor Anaphora Resolution System for Estonian. Tartu Ülikool, Arvutiteaduse instituut. 2008
- Kadri Kajaste. *Morfoloogilisest ühestamisest*. 2009

Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs 2009-2010

Eesmärgid

Projekti eesmärgiks on olemasolevale morfoloogilisele ühestajale ja pindsüntaktilisele analüsaatorile tuginedes luua:

1. Grammatikakorrektori tööversioon
2. Suulise keele süntaksianalüsaatori arendamine.
3. Murdetekstide süntaktiline analüüs
4. Interneti keele (uue meedia keele) süntaktiline analüüs
5. Õppijakeele süntaktiline analüüs
6. Sügavamate sõltuvusseoste tuvastamine