



Automatic Constraint Grammar shallow syntactic parsing of spoken Estonian

Kaili Müürisep Heli Uibo
Institute of Computer Science
University of Tartu



Outline

- Motivation
 - Corpus of Spoken Estonian
 - Morphological analysis and disambiguation
 - Constraint Grammar
 - Tag set
 - Modification of rules
 - Results
 - CG2Tree
 - Conclusions
-



Motivation

- Existing morphologically disambiguated corpus of **spoken** Estonian
 - Existing parser for **written** Estonian
 - Curiosity
-



Corpus of Spoken Estonian

- Started 1997 (Tiit Hennoste et al.)
- Open corpus, no upper limit
- Different types of spoken language: everyday and institutional conversations, spontaneous and planned speech, monologues and dialogues
- Max. authentic situations
- 700,000 transcribed words



Morphological analysis and disambiguation

- ESTMORF morphological analyzer and guesser, adapted for spoken language texts
(recognizes e.g. *kolmkend* = kolmkümmend = thirty)
- hand-corrected
- disambiguated manually



Input

K	####	
\$<s>		
muna	muna+0 // _S_ com sg nom //	egg
noh	noh+0 // _B_ //	well
see	see+0 // _P_ dem sg nom //	this
siia	siia+0 // _D_ //	here
asemele	asemele+0 // _D_ //	instead of
tuleks	tule+ks // _V_ main cond pres ps3 sg ps af #FinV #Intr //	should
leida	leid+a // _V_ main inf #NGP-P //	find
midagi	miski+dagi // _P_ indef sg part //	something
muud	muu+d // _P_ indef sg part //	else
ma	mina+0 // _P_ pers ps1 sg nom //	I
soovitaks	soovita+ks // _V_ main cond pres ps1 sg ps af //	suggest
hapukoort	hapu_koor+t // _S_ com sg part //	sour cream
\$.	. // _Z_ Fst //	
\$</s>		



Constraint Grammar Parser for Estonian

- Uses the first version of Constraint Grammar
- Designed for written language
- Tagset: SUBJ OBJ PRD ADVL +FMV -FMV +FCV -FCV P>
<P Q> <Q NN> <NN AN> <AN PN> <PN etc.
- Very shallow, dependency oriented
- Ca 1200 syntactic constraints and 50 clause boundary detection rules.



New syntactic labels

- New part-of-speech – special particles – *ahah*, *mhmh*, *hurraa*, *jess*, *ee*, *õõ*, *noh* etc. These are already marked by morphological analyzer.
- Parser annotates these with special label:
B - syntactically independent uninflected words
- T – unknown syntactic function, used both for word forms with no morphological information and for word forms with an unclear syntactic function.



Modification of rules

1. compile new rules for the sentence internal clause boundary detection
2. fix the syntactic constraints taking into account the specific features of the spoken language (slight modifications of less than 100 rules from 1200)



Sentence internal clause boundaries

- Parser considers speech turn in dialogues as a unit of analysis (sentence).
- Pauses are marked by punctuation marks – parser uses them for detecting clause boundaries
- Some particles and adverbs are used in the beginnings or ends of clauses



Modification of Rules

- We also had to inspect and revise all erroneous syntactic rules.
- In order to accomplish this task, we have manually compiled a syntactically annotated benchmark corpus of 2200 words.



Output

```
K      #####
$<s>
muna   muna+0 // _S_ com sg nom //  **CLB @SUBJ      ;; egg
noh    noh+0 // _B_ //  @B                          ;; well
see    see+0 // _P_ dem sg nom //  @<NN             ;; this
siia   siia+0 // _D_ //  @ADVL                       ;; here
asemele asemele+0 // _D_ //  @ADVL                   ;; instead of
tuleks tule+ks // _V_ main cond pres ps3 sg ps af #FinV #Intr //  @+FMV      ;; should
leida  leid+a // _V_ main inf #NGP-P //  @OBJ        ;; find
midagi miski+dagi // _P_ indef sg part //  @OBJ      ;; something
muud   muu+d // _P_ indef sg part //  @<NN          ;; other
ma     mina+0 // _P_ pers ps1 sg nom //  **CLB-C @SUBJ ;; I
soovitaks soovita+ks // _V_ main cond pres ps1 sg ps af //  @+FMV      ;; suggest
hapukoort hapu_koor+t // _S_ com sg part //  @OBJ    ;; sour cream
$.     . // _Z_ Fst //
$</s>
```



Results

- The word count in the corpus: 2194
 - Errors: 68
 - Recall: 96.9% (98.5%)
 - Precision: 89.5% (87.5%)
 - Unambiguity rate: 92.9% (89.5%)
-



Errors

1. inadequate inner clause boundary detection: 16
2. unknown tag: 12
3. postmodifying attribute: 5
4. adjective functioning as a noun: 9
5. heuristic rules: 3
6. earlier wrong analysis: 5
7. repetition: 3
8. other: 14



Example

selle taga on saad aru selline lähenemine

this behind is-SG3 understand-SG2 this approach

/this approach is used behind this as you understand/

- The subject tag has been removed from word form *lähenemine* since it can't co-exist with the verb 2nd person singular.



Repetitions

aga miks miks miks peab ...

but why why why must

Aga sa aga sa peaksid katsuma kompressida ...

but you but you should try to compress

See võtab noh mahutab rohkem

This takes noh accomodates more

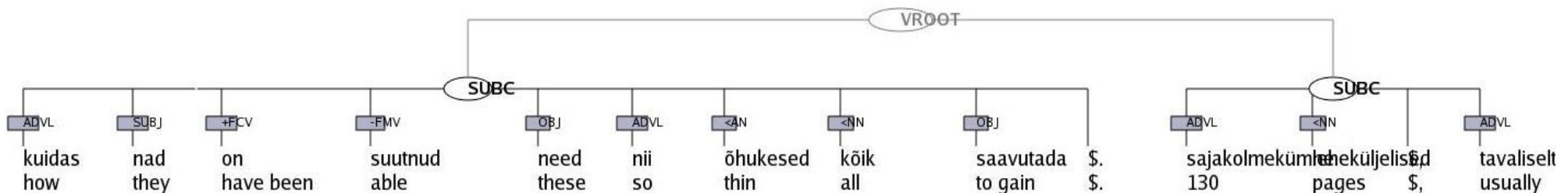


Spoken language specific annotation

- Unfortunately, we had to ignore the spoken language specific annotation (overlapping dialogue, speech acts etc), as we have not yet worked out the method, how to represent this information in the syntactic tree.

Cg2Tree

- The sample corpus was converted to Negra export format by a Perl program written by Kaarel Kaljurand
- Next, we imported the treebank to TigerSearch.
- The trees are very flat yet - the smallest group is a subclause. For tree deepening we might try to use the approach used for the semi-automatic creation of the VISL-treebank Arborest (<http://corp.hum.sdu.dk/arborest.html>).





Conclusions and plans for future

- Analysis of spoken language was not as complicated as we expected
- The generated tree should be deeper
- The tree should represent also spoken language specific information