Shallow Parsing of Spoken Estonian Using Constraint Grammar

Kaili Müürisep and Heli Uibo
Institute of Computer Science
University of Tartu, Estonia

Abstract

In this paper we describe how we have adapted the syntactic analyzer of written Estonian to the spoken language. The Constraint Grammar shallow syntactic parser (Müürisep et al. 2003) was used for the automatic syntactic analysis of the corpus of Estonian spoken language (Hennoste et al. 2000). To adapt the parser, the clause boundary detection rules as well as some syntactic constraints had to be changed. Two new syntactic tags were also introduced. In the paper the introduced changes are described and the achieved results are analyzed. The parser determined the syntactic label unambiguosly for 90% of the words in the text in average, using the manually morphologically disambiguated text as an input. The error rate was less than 3%.

1 Introduction

Manual syntactic annotation of spoken language corpora is a cumbersome and time-consuming task. In order to simplify the process, we have employed the Estonian Constraint Grammar parser (Müürisep et al. 2003) which was originally designed for the automatic analysis of the written language.

The creation of the Tartu University Corpus of Spoken Estonian (Hennoste et al. 2000) started in 1997. The corpus currently contains 700,000 running words of spoken dialogues and monologues. The corpus is transcribed by the transcription of conversational analysis (CA).

Some parts of the corpus are morphologically analyzed and disambiguated, some parts are annotated with labels of dialogue acts (Hennoste et al. 2003).

The corpus is open by its design, i.e., new texts can be freely added and no limits to the size of the corpus have been set. The intention is to collect

various types of oral speech, both everyday and institutional conversation, spontaneous and planned speech, monologues and dialogues, face-to-face interaction and media texts.

The texts that we have used in our experiments have been morphologically analyzed and disambiguated. ESTMORF morphological analyzer was used in a special guessing mode developed for morphological analysis of spoken language (Hennoste et al. 2002), as in the corpus of spoken language the words are transcribed as they are spoken (often differently from the correct orthographic form). For example, *kolmkend* is analyzed as *kolmkümmend* (thirty). However, some words have been analyzed or corrected manually.

To adapt the parser for the spoken language, we had to compile new rules for the sentence internal clause boundary detection and fix the syntactic constraints, taking into account the specific features of the spoken language.

The similar experiment has been described by Bick (1998) for Portuguese, adapting Constraint Grammar based tagger/parser for written Portuguese as a tool for annotating Brazilian urban speech corpus. According to Bick, the performance for tagging was good (error rate under 1%) but the syntactic error rate deteriorated (5%).

The following sections give an overview of the Estonian Constraint Grammar (EstCG) parser, describe the process of adapting rules, and discuss the results. It should be mentioned that we have not addressed morphological disambiguation (or tagging) problems in this paper.

2 Constraint Grammar parser of written Estonian

The EstCG parser was developed in 1996-2000 by T. Puolakainen and K. Müürisep.

The main idea of the Constraint Grammar (Karlsson et al. 1995) is that it determines the surface-level syntactic analysis of the text which has gone through prior morphological analysis. The process of syntactic analysis consists of three stages: morphological disambiguation, identification of clause boundaries, and identification of syntactic functions of words. Grammatical features of words are presented in the forms of tags which are attached to words. The tags indicate the inflectional and derivational properties of the word and the word class membership, the tags attached during the last stage of the analysis indicate its syntactic functions.

27 syntactic tags of EstCG represent syntactic functions of traditional Estonian grammar (Erelt et al. 1993), although there are some modifications considering the specialities of Constraint Grammar: CG annotates each word

with some syntactic label while linguistic grammar has a more general view, treating multiple words as units.

The syntax used in CG is word based, i.e., no hierarchical phrase structure is constructed. The phrasal heads are labelled as subjects, objects, adverbials or predicatives. The modifiers have tags that indicate the direction where the head of a phrase could be found but the modifiers and heads are not formally connected. The components of a verb chain are marked by five labels: finite or infinite auxiliary or main verb and a label for negation.

Determination of syntactic functions is implemented in two modules. First, the parser adds all possible function tags to each morphological reading, and after that, syntactic constraints remove incorrect tags in the current context.

CG consists of hand written rules which decide by checking the context whether an interpretation is correct or has to be removed.

A number of rules are clearly of a heuristic nature – the rule might not be 100 % true but its proficiency rate is very high, compared to the number of errors. Several rules have been compiled solely on the statistical information about the word order in the sentence.

The rules are grouped in such a way that the most reliable ones or those that cause least errors are in the main part of the grammar; the heuristic rules have been divided into groups based on their reliability.

The grammar consists of 1,240 morphological disambiguation rules, 47 clause boundary detection rules, 180 morphosyntactic mapping rules and 1,118 syntactic constraints.

The morphological disambiguation rules are thoroughly discussed in (Puolakainen 2001) and syntactic constraints in (Müürisep 2000).

As a result of tests, 86.6 % of words become morphologically unambiguous, and the error rate of the morphological disambiguator is 1.8 %.

The results of the full analysis show an ambiguity rate of 17 % (83 % of all word forms are unambiguous) and an error rate of 3.5 % (Müürisep et al. 2003).

A disambiguated and syntactically analyzed sentence is shown in Figure 1. Morphological description is placed between "//"-symbols, syntactic tags begin with the @-symbol. The direct translation is given after the #-symbol. The last word in the sentence remains ambiguous between adverbial and postmodifying attribute. The phrase *koht infootsingul* ('place on the information retrieval') has no meaning but the attribute tag can not be removed since the phrase with some other attribute in adessive case is quite usual, e.g., *koht laeval* – 'place on the ship'.

EstCG parser is based on the original Constraint Grammar framework but has been reimplemented. It has some influence from CG-2 (Tapanainen

```
$LA$
   ####
Dokumenditöötluses           # in the document processing
   dokumendi_töötlus+s //_S_ com sg in cap // **CLB @ADVL
on                           # is
   ole+0 //_V_ main indic pres ps3 sg ps af Intr // @+FMV
oluline                      # important
   olu=line+0 //_A_ pos sg nom //  @AN>
koht                         # place
   koht+0 //_S_ com sg nom  //  @SUBJ
infootsingul                 # on the information retrieval
   info_otsing+l //_S_ com sg ad  //  @ADVL @<NN
$.
   . //_Z_ Fst //
$LL$
   ####
```

Figure 1: Syntactically analyzed sentence - *Information retrieval has an important place (role) in the document processing.*

1996), like options for enhanced context addressing and for morphological disambiguation after the phase of determination of syntactic functions. Our parser also enables rules for clause boundary detection and these rules are exploited in EstCG.

3 Modifications in tag set and rules

3.1 New labels for spoken language corpus

We have adapted two additional tags: @B – particle; @T – unknown syntactic function, used both for word forms with no morphological information and for word forms with an unclear syntactic function.

Particles occur very frequently in the spoken language. Estonian particles are considered as an independent part of speech. Most of the particles are indeclinables with their own root. From the phonetic viewpoint, some of the particles are the words existing in the written language (*siis*, *jah*), some are their pronunciation variants (*sis*, *kule*), some are the phoneme combinations which are balancing on the borderline of a word and a sonification (*ööäk*, *phtüi*, *mhmh*). Syntactically, these units may form a whole speech act, thus a whole syntactic unit (*mhmh*). If the particles belong to a longer intonational unit, e.g., sentence, then they do not belong to the grammatical structure of the sentence. They do not have any syntactic role in the sentence, but act as

free modifiers of the sentence as a whole. Semantically, the particles have (almost) no meaning (Hennoste 2002).

There are quite many ungrammatical or unfinished sentences in the spoken speech, also the words may be incomplete or uttered in the noisy environment, and therefore not properly transcribed and morphologically analyzed as unknown words. The correct determination of all syntactic functions in ungrammatical sentences may be impossible even for a human annotator. These words are labelled as words with unknown syntactic function - @T. The sample sentence in the Figure 2 is unfinished and therefore two last word forms have unclear syntactic functions.

```
$<s>
    ####
kui                                                 #if
    kui+0 //_J_ sub //   **CLB @J
sa                                                  #you
    sina+0 //_P_ pers ps2 sg nom //   @T
võtame                                              #take
    võt+me //_V_ main indic pres ps1 pl ps af NGP-P // @+FMV
mingisuguse                                         #some
    mingi_sugune+0 //_P_ indef sg gen //   @NN>
asja                                                #thing
    asi+0 //_S_ com sg gen //   @OBJ
kuigi                                               #although
    kuigi+0 //_J_ sub //   **CLB @J
seda                                                #this
    see+da //_P_ dem sg part //   @T
see                                                 #this
    see+0 //_P_ dem sg nom //   @T
$</s>
```

Figure 2: Example of the usage of the label for unknown syntactic function

3.2 Syntactic window

In the case of the spoken language, the segmentation of the speech into sentences is a complex task. In the written language, the sentence is considered as one syntactic unit or a syntactic window, which defines the scope of the context used for parsing. The sentence of the written language has a punctuation mark-up added by the author of the sentence. In the spoken language, the falling and rising intonations may be used as the delimiters of the sentences.

The input text of the parser was segmented to utterances by the transcription marks using the following scheme: 1. The unit we call as utterance ends

with a clearly detectable falling intonation that indicates the termination of the unit. It is marked by the full stop. 2. Each utterance can be divided into intonational parts separated by less falling intonations than in the end of the utterance. The separating intonation that marks the boundary (but does not terminate the unit) is marked by the comma. 3. In addition, the unit ending with a rising intonation is marked by the question mark.

When looking the transcription mark-up more closely it became clear that the segmentation of the text into utterances was not precise enough — the intonation may often fall inside the syntactic unit also (see Example 1).

(1) ma tõstan kartulid ära . ja sousti ka .
I shall lift potatoes aside and sauce also

This was the reason why we decided to consider a turn in the dialogue as a syntactic window and treat the units separated by full stops as coordinated clauses.

The style of the determination of the sentence boundary depends on the type of the text, and the usage of full stops as sentence delimiters is justified in the case of monologues.

On the other hand, there are many examples of situations where dialogue turn divides the sentence into two separate parts, e.g., Example 2. The word forms *auto* (car) and *saada* (to get) have been labelled as unknown since it is impossible to determine their syntactic function inside the turn.

(2) A: see oli kõik ee ausatel eesmärkidel et perele auto
this was all ee honest goals that family car

B: ei no loomulikult
of course

A: saada. aga lissalt mai saand seda autot
get but simply I didn't get this car

3.3 Clause boundary detection rules

The clause boundaries in written text are determined by rules based on conjunctions, punctuation marks and verbs. The basic rule runs as follows:

> If a word is preceded by a punctuation mark and/or the word itself is a conjunction and in the right and left contexts there is a conjugable form of a verb then the word is the first word of a clause.

The conditions of this rule may vary. For example, the comma and coordinating conjunctions *ja, ning, või, ega, ehk* may separate not only clauses but also coordinated phrases or words. Therefore, we use two tags for annotating clause boundaries — CLB marks absolutely clear and sure clause boundaries, and CLB-C is used in suspicious cases. There are 47 clause boundary detection rules in EstCG, a number of them for very specific cases.

These clause boundary detection rules have been thoroughly revised since the meaning and usage of punctuation marks have been changed.

The punctuation marks (that describe the intonation) have been taken into account during the automatic inner clause boundary detection; also, a special attention has been paid to particles characterizing the spoken language (*noh*, pause fillers *aa, ee, öö*). For example, the full stop (detectable falling intonation) is used as a sure delimiter of clause. Particles are used as delimiters if there are finite main verbs in both left and right contexts.

The rules try to discover false starts and mark these by clause boundary tags but this is possible only if there is a verb in the false start phrase, e.g.,

(3) mul (CLB) on kassetil (CLB-C) oleks
I-SG-AD be-SG3 tape-SG-AD be-COND
ruumipuudus tekkinud
lack_of_space-SG-NOM arise-PCP
'I have CLB-C there would be no space on the tape'

The main source of errors are the endings of embedded clauses which are hard to detect, e.g.,

(4) kuna (CLB) ta tundus mulle esmakuulamisel või noh algul
as he seemed me listening-first-time or noh first
kui (CLB) ma kuulasin (*) kuidagi liiga afishlik või noh noh
when I listened somehow too posterized or noh noh
äraleierdatud
hackneyed
'As he seemed me too posterized or hackneyed when I listened for the first time'

This clause boundary would have been marked with a comma in a written text.

The grammar for the spoken language consists of 22 clause boundary detection rules, which is remarkably lesser than for the written language. This may be explained by the fact that the grammar for the written language is adapted for analyzing legal texts and the rules consider the different punctuation styles for laws.

3.4 Modification of the syntactic constraints

As already mentioned in this paper, the original grammar for thewritten language consists of 1118 rules. These rules were applied to the corpus of 2200 words (everyday conversation) which was manually syntactically annotated. The erroneous rules were modified or complemented. Mostly the contextual conditions were changed. The rules of EstCG may be applied in three different modes: a) the scope of contextual conditions is a whole sentence; b) the scope of contextual conditions is inside the sure clause boundary markers (CLB); c) the scope of contextual conditions is inside all clause boundary markers (both CLB and CLB-C).

The main source of errors was the misuse of the context: the rules took into account the far context situating outside the clause. In order to fix it only minor changes were made: the sign of the mode of contextual conditions was replaced as demonstrated in Example 5. This rule removes the subject tag if the word form is in partitive case and there is a first person singular verb in the left context. The first rule is the original, and the second is the modification.

(5)
```
(@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB)
(@w =s0 (@SUBJ) (0 Par)(*-1C Sg1) **CLB-C)
```

The rules which make use of punctuation marks were also inspected thoroughly and new conditions were added for the sentences without commas.

The usage of vocabulary is different in the spoken language. We had to consider, for example, that relative pronoun *mis* (what) may be used in the interrogative sentence instead of the question word *kas* (whether) or in the comparisons instead of *nagu* and *kui* (than).

The modification of rules was finished when the error rate descended from 7.5% to 3%.

4 Evaluation

4.1 Description of the test corpus

The new test coprus[1] of 2543 words was used for the evaluation of the performance of the parser. This corpus was manually syntactically annotated in order to compare the analyses automatically. The test corpus consists of everyday conversations, representing both longer and shorter dialogue sentences. The corpus was compiled semi-automatically. One human annotator read the automatic analysis and fixed the errors caused by the parser.

[1] Available at http://www.ut.ee/~kaili/Korpus/Spoken

The authors admit that this type of evaluation has its shortcomings:

1. The corpus has a small size and is not representative.
2. The human annotator may not notice all the errors in the automatically analyzed corpus.
3. Several persons should annotate the corpus in parallel, in order to detect all the errors and follow all the tagging instructions.

4.2 Results

The output of the parser was compared with manually annotated corpus and the following results have been achieved (the results for parsing the written language are enclosed in parentheses):

- the word count in the corpus: 2543;
- recall (the ratio of the number of correctly assigned syntactic tags to the number of all correct tags): 97.3% (98.5%);
- precision (the ratio of the number of correctly assigned syntactic tags to the number of all assigned syntactic tags) : 89.2% (87.5%);
- unambiguity rate: 91.5% (89.5%).

4.3 Types of errors

The errors may be classified as following (the percentage is calculated on the basis of the parser performance on the training corpus).
1. Errors caused by inadequate inner clause boundary detection – 17%. (See Example 6.)

(6) selle taga on saad aru selline
this-GEN behind-POST be-SG3 understand-SG2 this-NOM
lähenemine
approach-NOM
'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. The rules are not able to detect the end of the embedded clause.
2. Unknown syntactic function – 19%. There are a lot of incomplete, elliptical or ungrammatical sentences in the spoken language, therefore even

the human annotator is unable to analyze all the words in the sentence. The word forms with unknown syntactic functions are labelled with @T-tag. On the other hand, it is complicated (or even impossible) to write rules for the parsing grammar which declare on the basis of context that some word form is ungrammatical in this context. EstCG parser adds all possible syntactic labels to every word form on the basis of morphological information and context; and by applying syntactic constraints, removes the labels not fitting into the context one by one. If the sentence is incomplete then the parser has not enough contextual information and the word form may remain ambiguous. (See Example 7, the correct labels are in the square brackets, the labels determined by the parser begin with @-symbol.)

On the other hand, the parser may attempt to remove all labels due to erroneous context. Fortunately, one of the design principles of Constraint Grammar is to never remove the last reading or label. The Example 8 points out that the last label (premodifying attribute) may be rather odd in the context.

(7) A: et [J] @J nad [T] @SUBJ @OBJ
that-J they-PL-NOM

B: mhmh [B] @B
mhmh (yes)

A: sobivad [FMV] @FMV kätte [ADVL] @ADVL
fit-PL-3PRS hand-ILLAT

'A: that they B: yes A: fit to hand'

(8) pöidla [T] @NN> ja [J] @J kõik [OBJ] @OBJ on [FMV] @FMV
thumb-GEN and all be-SG-3PRS
tehtud [FMV] @FMV
do-PCP-IMP
'thumb and all is done'

3. Adjective functioning as a noun – 12%. Adjectives can act as subjects or objects only in elliptical sentences. Typically these cases are very irregular and are not covered by EstCG rules.

(9) ma @SUBJ ostaks @FMV selle @NN>
I-NOM buy-COND this-SG-GEN
maasikamaitselise @OBJ
strawberry_flavoured-SG-GEN
'I would buy this strawberry flavoured [one]'

4. Earlier wrong analysis – 5%. One of the preceding rules has caused an error and this wrongly analyzed word form leads to new errors.
5. Repetitions – 3%. If a subject is repeated then the principle of uniqueness has been violated.

(10) noh [B] @B se [SUBJ] @NN> see [SUBJ] @SUBJ
noh this-SG-NOM this-SG-NOM
on [FMV] @FMV tähtis [PRD] @PRD
be-SG-3PRS important-SG-NOM
'you know this this is important'

6. Other errors. These may occur also in written language texts.

4.4 Ambiguities

When we compared the remained ambiguity classes of the spoken and written language then it turned out that their structure is different. The dominating ambiguity class in the automatically analyzed corpus of the written language was the ambiguity between adverbial and postmodifying attribute. The ambiguities between object and premodifying attribute, adverbial and premodifying attribute and subject and object occurred about three times less frequently. The leading ambiguity class in the spoken language is formed by infinite verbs which are tagged both as adverbials and parts of predicate chain. The adverbial and subject, adverbial and postmodifying attribute, and subject and object are next in the ranking.

5 Conclusion and plans for future

The adaption of the parser of the written language for the spoken language turned out to be easier task than expected. The efficient detection of clause boundaries became the key issue for successful automatic analysis, while syntactic constraints required only minimal modification. Quite surprisingly, the performance of the parser for the spoken language exceeded its original performance for the written language. There are two reasons for that. 1) There are lot of particles in the spoken language that involve only trivial syntactic analysis. 2) The adverbial is defined as the *remainder* class in EstCG that accommodates all word forms excluded from other word classes. The adverbials, especially the adverbs as adverbials, are quite numerous in the spoken language sentences, and their analysis is an easy task for the parser. Therefore, the handling of these word classes increases the correctness and precision of the parser.

The evaluation revealed the suitability of the Constraint Grammar formalism for analyzing the spoken language: 1) since no exact link exists between the modifier and its head, wrongly agreeing word forms may be still correctly analyzed; 2) words that can't be analyzed are left ambiguous which lessens the number of errors.

On the other hand, one should be cautious when assessing how easy it would be to adapt the EstCG morphological disambiguator for the spoken language. E. Bick (1998) reached the conclusion that the modification of disambiguation rules of Portuguese CG provides better results than the modification of syntactic rules, since disambiguation involves a smaller context. However, this conclusion might not hold for Estonian which is known for its free word order and rich morphology. Hopefully further experiments will provide a deeper insight into this issue.

One of the shortcomings of the grammar we have created is the lack of tagging for the spoken language specific constructs like repetitions, false starts and self corrections. The authors however believe that this sort of tagging does not logically fit into the existing framework, but should be carried out by some other tool set or formalism (the presence of such tagging would ease the determination of syntactic functions).

We have not yet worked out the method how to represent this syntactic information in the syntactic tree. For treebank generation we might try to use the approach used for the semi-automatic creation of the VISL-treebank Arborest[2] (Bick et al. 2004).

References

Bick, E. (1998). Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese. In *Proceedings of the 17th Scandinavian Conference of Linguistics*, Odense.

Bick, E., H. Uibo, and K. Müürisep (2004). Arborest - a VISL-Style Treebank Derived from Estonian Constraint Grammar Corpus. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*, Tübingen.

Erelt, M., R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, and S. Vare (1993). *Eesti keele grammatika. II Süntaks*. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Hennoste, T. (2002). Suulise kõne uurimine ja sõnaliigi probleemid. In *Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse*

[2] http://corp.hum.sdu.dk/arborest.html

õppetooli toimetised 4, pp. 56–73. Tartu: Tartu ülikooli kirjastus.

Hennoste, T., M. Koit, A. Rääbis, K. Strandson, M. Valdisoo, and E. Vutt (2003). Developing a Typology of Dialogue Acts: Tagging Estonian Dialogue Corpus. In I. Kruijff-Korbayová and C. Kosny (Eds.), *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, pp. 181–182.

Hennoste, T., L. Lindström, O. Gerassimenko, A. Jansons, A. Rääbis, K. Strandson, P. Toomet, and R. Vellerind (2002). Suuline kõne ja morfoloogiaanalüsaator. In R. Pajusalu and T. Hennoste (Eds.), *Tähendusepüüdja. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 3.*, pp. 161–171. Tartu: Tartu ülikooli kirjastus.

Hennoste, T., L. Lindström, A. Rääbis, P. Toomet, and R. Vellerind (2000). Tartu University Corpus of Spoken Estonian. In T. Seilenthal, A. Nurk, and T. Palo (Eds.), *Congressus Nonus Internationalis Fenno-Ugristarum 7.-13. 8. 2000. Pars IV. Dissertationes sectionum: Linguistica I*, Tartu, pp. 345–351.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila (1995). *Constraint Grammar: a Language Independent System for Parsing Unrestricted Text*. Berlin and New York: Mouton de Gruyter.

Müürisep, K. (2000). *Eesti keele arvutigrammatika: süntaks*. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu: Tartu ülikooli kirjastus.

Müürisep, K., T. Puolakainen, K. Muischnek, M. Koit, T. Roosmaa, and H. Uibo (2003). A New Language for Constraint Grammar: Estonian. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 304–310.

Puolakainen, T. (2001). *Eesti keele arvutigrammatika: morfoloogiline ühestamine*. Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu: Tartu ülikooli kirjastus.

Tapanainen, P. (1996). *The Constraint Grammar Parser CG-2*. Publications of the Department of General Linguistics, No. 27. Helsinki: University of Helsinki.