Determination of Syntactic Functions in Estonian Constraint Grammar

Kaili Müürisep
Institute of Computer Science
University of Tartu
Liivi 2, 50409 Tartu
ESTONIA
kaili@ut.ee

Abstract

This article describes the current state of syntactic analysis of Estonian using Constraint Grammar. Constraint Grammar framework divides parsing into two different modules: morphological disambiguation and determination of syntactic functions. This article focuses on the last module in detail. If the morphological disambiguator achieves the precision more than 85% and error rate is smaller than 2% then 80-88% of words becomes syntactically unambiguous. The error rate of parser is 1-4% depending on the ambiguity rate of input. The main goal of this work is to elaborate an efficient parser for Estonian and annotate the Corpus of Estonian Written Texts syntactically. It is the first attempt to write a parser for Estonian.

1 Introduction

The main idea of the Constraint Grammar (Karlsson, 1990) 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. This article focuses on the last module in detail. 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. The underlying principle in determining both the morphological interpretation and the syntactic functions is the same: first all the possible labels are attached to words and then the ones that do not fit the context are removed by applying special rules or constraints. Constraint Grammar consists of hand written rules which by checking the context decide whether an interpretation is correct or has to be removed.

Constraint Grammar seemed to suit best for the analysis of Estonian texts because its mechanism is simple and easily implementable, it can be well adapted for the Estonian language, it is at the same time sufficiently reliable (robust) and the resulting syntactic analysis that the Grammar gives suits various practical applications.

2 Syntactic Analysis of Estonian

The Estonian language is a Finno-Ugric language and has got a rich structure of declensional and conjugational forms. The order of sentence constituents in Estonian is relatively free and influenced more by semantic and pragmatic factors.

For morphological analysis of Estonian, we use the morphological analyser ESTMORF (Kaalep, 1997) that assigns adequate morphological descriptions to about 98% of tokens in a text. Morphologically analysed text is disambiguated by Constraint Grammar disambiguator of Estonian. The development of disambiguator is in process but 85-90% of words become morphologically unambiguous and the error rate of this disambiguator is less than 2% (Puolakainen, 1998).

All the syntactic information is given by syntactic tags in constraint grammar framework. The syntactic tags of Estonian Constraint Grammar (ESTCG) are derived from tag set of English Constraint Grammar (ENGCG) (Voutilainen et al., 1992) with some modifications considering the specialities of Estonian. These tags are attached to words by 175 morphosyntactic mapping rules. After this step of parsing there are approximately 3.8 tags per word.

After the mapping operation syntactic constraints are applied. ESTCG contains 800 syntactic constraints. In fact, nearly half of them treat

the attributes. It can be explained by the fact that there are 12 types of attributes in ESTCG and the attribute tags are also added to almost every word in sentence (except finite verbs and conjunctions).

3 Results

To evaluate the performance of parser I use two types of corpora. Training corpus is used for formulating rules and preliminary testing. After testing I improve rules so that most errors will be fixed next time. Benchmark corpus is used only for evaluating parser. Both types of corpora consist of fiction texts. The training corpus contains 4 texts of 2000 words from different Estonian writers. Benchmark corpus consists of 2000 word. I used these corpora in two experiments. In the first experiment (experiment A) I tested only the syntactic function detecting part of grammar and I supposed that the input text is ideally morphologically analysed and disambiguated, this means that all words are morphologically correct and unambiguous. For this experiment both corpora were manually morphologically disambiguated. In the second experiment (experiment B) I used the same corpora but they were disambiguated automatically. In this case the disambiguator made 2% errors and left 13% of words ambiguous, 1% of words were unknown for morphological analyser.

The precision and recall of ESTCG parser are shown in table 1.

Table 1. Recall and precision.

Corpus	Recall	Precision
A Training	99,12%	83,76%
A Benchmark	98,12%	85,00%
B Training	95,76%	74,34%
B Benchmark	96,58%	76,52%

The big number of errors in B experiment can be explained by the fact that I wrote preliminary grammar rules using only manually disambiguated corpora and the work on correcting rules using more ambiguous input is still in process. As I mentioned before the input was ambiguous and erroneous in this experiment and this caused error rate of 3%.

The errors in manually disambiguated corpora are mostly caused by ellipsis, some errors occurred during determination of apposition and the third biggest group of errors exists in sentences there one clause divides the other into two parts.

In experiment A, 86-88% of words become syntactically unambiguous, and in experiment B, the corresponding numbers are 80-82%. In both experiment less than 0,5% of words have 5-6 syntactic tags.

It is very difficult to distinguish adverbial attributes and adverbials. Approximately 6% of analysed words have both labels. This is almost the same problem as PP-attachment in English but additionally it is possible to use both premodifying and postmodifying adverbial attributes in Estonian. Of course the PP-attachment problem is also existent. The other hard problem is the distinction of genitive attributes and objects. If two or more nouns in genitive case are situated side by side then these words remain usually ambiguous, e.g. ... *siis vabastab kohus tema vara hooldaja järelevalve alt.* / ... then free-SG3 court-NOM he-GEN property-GEN trustee-GEN supervision-GEN from-POSTP / '... then the court frees his property from the supervision of trustee.'

4 Conclusions

In this paper I described my work on the syntactic part of Estonian Constraint Grammar parser. The error rate of parser is 1-4% depending on ambiguity rate of input. 80-88% of words become syntactically unambiguous.

The most exhaustive Constraint Grammar is written for English. Timo Järvinen, the author of syntactic part of ENGCG, reported that the error rate is 2 - 2,5% and ambiguity rate ca 15% (Järvinen, 1994). Of course the Estonian and English are too different languages and the comparison of performance of parsers do not help to draw any fundamental conclusions. But I really hope that the Estonian parser achieves nearly the same performance very soon. The further work will focus on decreasing the error rate and using statistical analysis for generating new rules.

References

Timo Järvinen. 1994. Annotating 200 Million Words: The Bank of English Project. In *Proceedings of COLING-94.* Vol. 1, 565-568, Kyoto.

Heiki-Jaan Kaalep. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on its Development. *Computers and Humanities*, 31(2):115-133.

Fred Karlsson. 1990. Constraint Grammar as a framework for parsing running text. *Proceedings of COLING-90.* Vol. 3, 168-173, Helsinki.

Tiina Puolakainen. 1998. Developing Constraint Grammar for Morphological Disambiguation of Estonian. *Proceedings of DIALOGUE'98.* Vol. 2, 626-630, Kazan.

Atro Voutilainen, Juha Heikkilä and Arto Anttila. 1992. *Constraint Grammar of English. A Performance Oriented Introduction.* Publications 21, Department of General Linguistics, University of Helsinki.