PARSING ESTONIAN WITH CONSTRAINT GRAMMAR

Kaili Müürisep
Institute of Cybernetics at Tallinn Technical University
Estonia
kaili@phon.ioc.ee

Abstract

This paper describes the current state of syntactic analysis of Estonian using Constraint Grammar, focusing mainly on the determination of syntactic functions. Constraint Grammar of Estonian was written in 1996-2000 at the University of Tartu. The author has developed its syntactic part.

1. Introduction

The work with syntactic analyser of Estonian started five years ago at the University of Tartu. As a basis of the analyser we are using a formalism called Constraint Grammar, which is developed at the University of Helsinki by Fred Karlsson and co-workers (Karlsson et al, 1995).

The main idea of the Constraint Grammar 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.

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 called constraints. Constraint Grammar consists of hand written rules, which by checking the context decide whether an interpretation is correct or has to be removed.

2. Preceding steps

For morphological analysis of Estonian, we use the morphological analyser ESTMORF (Kaalep, 1997) that assigns adequate morphological descriptions to about 99% of tokens in a text. In Estonian fiction texts about 45% of morphologically analysed word-forms have more than one reading.

Morphologically analysed text is disambiguated by Constraint Grammar disambiguator (Puolakainen, 1998). The development of the disambiguator is in process but 85-90% of words become morphologically unambiguous and the error rate of this disambiguator is less than 2%. The disambiguating grammar consists of more than 1200 hand written rules, almost half of them treat concrete word forms (e.g. '*on*' - verb *be* in simple present 3rd person singular or plural), the others cover broader ambiguity classes. The difficult problem is the choice between the readings of a noun in nominative, genitive, partitive or short illative (aditive) case. The other sources of

errors and ambiguities are participles and readings of adposition, adverb and noun of some word-forms.

3. Determination of 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 every word with some syntactic label while linguistic grammar has a more general view treating multiple words as units. The syntax used in Constraint Grammar is word based, this means that 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 phrase could be found but the modifiers and heads are not formally connected. The verb chain is 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.

Syntactic tags are added to words by 180 morphosyntactic mapping rules. These rules describe which combination of syntactic tags should to be attached to the current morphological reading. For example, a noun in nominative case can be a subject, an object, a predicative, a premodifying or postmodifying attribute or an adverbial. In this stage of parsing at least one syntactic tag is assigned to every word but usually many more (approximately 3.8 tags per word in the case of Estonian).

After the mapping operation syntactic constraints are applied. ESTCG contains 1118 syntactic constraints. The rules were devised using training corpus of 20,000 words. Most of these rules have linguistic background, this means that they are generated using grammar books and author's personal linguistic intuition. Only some of them are compiled using statistical information about word order tendencies. As known, any natural language tends to have somehow irregular nature - it is very difficult (if not impossible) to describe a language with fixed rules. So ca 20% of syntactic constraints in ESTCG are heuristic rules - they are not 100% true but ease to solve some complicated ambiguity classes. ESTCG heuristic rules help to raise unambiguity rate from 79% to 91%, reducing correctness from 99.46% to 99.24% (results from training corpus). Attempts are made to devise rules, which will remove as few correct interpretations as possible and so result in as error-free analyses as feasible. The syntactic part of Estonian Constraint Grammar is fully documented in author's Ph.D. thesis (Müürisep, 2000).

A disambiguated and syntactically analysed sentence is shown in fig. 1. Morphological description is between "//"-symbols, syntactic tag begins with @-symbol. The direct translation is given after #-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't be removed since the phrase with some other attribute in adessive case is quite usual, e.g. *koht laeval - 'place on the ship'*.

```
$LA$
   ####
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 #FinV #Intr // @+FMV
oluline                                               # important
   olu=line+0 //_A_ pos sg nom  #line //  @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 analysed sentence - *'Information retrieval has an important place (role) in the document processing'*.

ESTCG parser is based on original Constraint Grammar framework but has been re-implemented by us. It has some influence from CG-2 (Tapanainen, 1996), like possibilities for enhanced context addressing and doing morphological disambiguation after the phase of determination of syntactic functions. Our parser enables also rules for clause boundary detection and these rules are in use in ESTCG grammar.

4. Evaluation

Two types of texts were used to evaluate the performance of the syntactic analyser. If the morphological disambiguation has been made manually, i.e. the input text was unerroneous, the recall (the ratio of number of correct assigned syntactic tags to the number of all correct tags) was 98.5% and the precision (the ratio of number of correct assigned syntactic tags to the number of all assigned syntactic tags) was 87.5%.

If the prior analysis of the same text was made automatically (in this case the disambiguator made 2% errors and left 13% of words ambiguous, 1% of words were unknown for morphological analyser) the recall was 96.5% and precision 78%. In the first text 86-91% of words became syntactically unambiguous, and in the second one, the corresponding numbers were 81-84%. The benchmark corpus consists of 10,000 words and has not been used during the rule generation process.

The errors in manually disambiguated corpora are mostly caused by ellipsis, some errors occurred during determination of apposition and the third biggest group of error exists in sentences there one clause divides the other into two parts. There was only one error due to morphological ambiguity in the second test; all the other additional errors were caused by the faults from earlier steps of analysis.

In spite of all efforts some words still remain ambiguous. For example, it is very difficult to distinguish adverbial attributes from the adverbials (see example above). 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 complicated problem is the distinction of genitive attributes and objects, which are followed by any other noun, e.g.

(1) Ta asetas mantli (gen @OBJ @NN>) tooli (gen @OBJ @NN>) seljatoele (@ADVL @<NN).

He put coat-GEN chair-GEN back-ALLAT.

'He put the coat onto the back of a chair.'

To make things even worse the morphological disambiguator often fails to solve the morphological ambiguities in the same position: noun in genitive case is frequently ambiguous between nominative or partitive case.

So the most difficult problem in the Estonian language appear to be determining the borders of noun phrases. It is often hard to decide which adjacent nouns belong to a common noun phrase, and which form separate noun phrases.

5. Further plans

Although the grammar is already effective enough to be used in practical applications, the authors of the grammar see the ways for further improvements.

We should increase the lexicon: in addition to valency information of verbs, the precise description of quantifiers and some type of adverbials is also needed. The presence of the lexicon of phrasal verbs is also essential.

We should increase the size of training and benchmark corpora and include new types of texts. This would enable experiments with statistical methods, which might be quite fruitful.

ESTCG parser is used as a part of the noun phrase parser and it is also included in experimental automatic summary generation software. We are looking for new application areas of the parser; the work of adjusting it for the needs of text-to-speech synthesiser is in progress.

6. Conclusion

This paper presents the application of the Constraint Grammar formalism to Estonian. Although this was the first attempt to write a computational grammar for Estonian, the achieved results show that the Constraint Grammar is flexible enough to use this framework for syntactic analysis of morphologically rich languages with relatively free word order like Estonian.

References

Erelt, Mati, R. Kasik, H. Metslang, H. Rajandi, K. Ross, H. Saari, K. Tael, S. Vare, 1993. *Eesti keele grammatika II.* Tallinn: ETA Eesti Keele Instituut

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

Karlsson, Fred, Arto Anttila, Juha Heikkilä, Atro Voutilainen, 1995. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Müürisep, Kaili, 2000. *Eesti keele arvutigrammatika: süntaks*. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.

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

Tapanainen, Pasi, 1996. *The Constraint Grammar Parser CG-2*. Publications of the Department ofGeneral Linguistics, University of Helsinki, No. 27.