A New Language for Constraint Grammar: Estonian*

Kaili Müürisep
Institute of Cybernetics
Tallinn Technical University
12168 Tallinn, Estonia
kaili@phon.ioc.ee

Tiina Puolakainen
Institute of Estonian Language
10136 Tallinn, Estonia
tiinap@eki.ee

Kadri Muischnek
Department of Estonian and
Finno-Ugric Linguistics
University of Tartu
50409 Tartu, Estonia
kmuis@ut.ee

Mare Koit and **Tiit Roosmaa** and **Heli Uibo**
Institute of Computer Science
University of Tartu
50409 Tartu, Estonia
{koit,roosmaa,heli_u}@ut.ee

Abstract

The Constraint Grammar of Estonian presented in the paper is the first attempt in automatic syntactic analysis of Estonian. The grammar consists of 1,240 morphological disambiguation rules, 47 clause boundary detection rules, 180 morphosyntactic mapping rules and 1,118 syntactic constraints. The rules have been devised using a training corpus of 20,300 words and have been tested on a benchmark corpus of 10,000 words. As the 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 demonstrate the ambiguity rate of 83% and error rate of 3.5%.

1 Introduction

The Estonian language is a Finno-Ugric language that has a rich structure of declensional and conjugational forms, and also a relatively free order of sentence constituents. In these respects it differs considerably from English. There are 14 cases in Estonian, but due to the free word order it is difficult to determine the syntactic functions of these cases. Furthermore, there is no grammatical gender. The person agreement (1st, 2nd, and 3rd person in singular and plural) is common in finite verbs in all forms and tenses. The majority of grammatical categories are implemented by means of morphology.

Estonian is characterised by a wide extent and variety of grammatical homonymy that makes the automatic analysis of Estonian a difficult task. In the case of English, the main difficulty lies in determining the correct part of speech. The same problem exists also in Estonian, but the number of choices is much greater due to the richness of forms.

The ratio of ambiguous words varies greatly from language to language: for example, in English, Swedish, and Finnish the ratio of words with multiple morphological interpretation is 40%, over 60%, and 11%, respectively (Karlsson *et al.* 95). In Estonian literary texts more than 45% of words are ambiguous. Estonian, unlike Germanic languages, is not subject-centered. There are a number of non-elliptical sentences in Estonian with no subject.

Before we started our project, an automatic morphological analyzer for Estonian had already been created (Kaalep 96). It was our task to elaborate a grammar suitable for the automatic syntactic analysis of Estonian, and to compile the program for the syntactic analysis. To accomplish this task, we had to choose a suitable grammar model for Estonian, and analyze the available Estonian texts, primarily from the Corpus of Written Estonian Texts (Hennoste *et al.* 98), in order to use the established regularities for wording the rules and writing the syntactic parser.

Our grammar has been composed on the formalism of the Constraint Grammar (Karlsson *et al.* 95). The main idea of the Constraint Grammar (CG) 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 work was supported by the Estonian Science Foundation under the grants No. 3314 and No. 4605.

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.

The Constraint Grammar parser of Estonian exists as two separate programs: the morphological disambiguator (Puolakainen 01) and the syntactic analyzer in a narrower sense (Müürisep 00). The basic differences of our grammar from the standard one are the following:

- the assumed clause boundaries are also used;
- the referenced context conditions can be represented in two ways: the appropriate cohort is searched either up to the very end of the possible context, or it is searched up to the first appropriate elements/filler;
- it is possible to remove morphological interpretations during the syntactic analysis.

The next sections of the paper will provide an overview of the Constraint Grammar of Estonian (EstCG) and problems that cropped up in its creation. We think that other researches who set the aim of elaborating an automatic syntactic analysis of a language can benefit from our experience, especially if the language is different from English, possesses rich morphology and/or free word order.

2 Motives for Selecting Constraint Grammar Formalism

According to Constraint Grammar, after the morphological analysis of a sentence the following steps are performed: morphological disambiguation, determination of sentence-internal clause boundaries, adding of syntactic tags, and finally, syntactic disambiguation.

As an example, let us consider the morphologically analyzed Estonian sentence "Aknas kustus tuli" (*The light went out in the window*):

```
Aknas (window)
   aken+s //_S_ com sg in //
kustus (go out)
   kustu+s //_V_ main indic impf ps3 sg ps af #Intr //
tuli (light)
   tule+i //_V_ main indic impf ps3 sg ps af #Intr //
(came)
   tuli+0 //_S_ com sg nom // (the light)
$.
```

Let us note that the word forms in this sentence can be sequenced in $3! = 6$ different ways, and all the resulting sentences will be correct and understandable for a native speaker of Estonian due to the free word order.

In analyzing this sentence, the correct interpretation of the word form *tuli* is found by applying the following constraint: remove the finite form of the verb from the cohort (in the present case verb tule+i //_V_ main indic impf ps3 sg ps af #Intr //) if a given word is immediately preceded by a finite form of verb which is the only interpretation of the word form (in the present case kustu+s //_V_ main indic impf ps3 sg ps af #Intr //).

After having added the syntactic tags we have the sentence in the following form:

```
Aknas
   aken+s //_S_ com sg in **CLB // @ADVL @<NN
@NN>
kustus
   kustu+s //_V_ main indic impf ps3 sg ps af #Intr //
@+FMV
tuli
   tuli+0 //_S_ com sg nom // @SUBJ @OBJ @ADVL
@NN> @<NN
$.
```

In this sentence, the noun *tuli* in the nominative case singular may be either the subject (@SUBJ), the object (@OBJ), adverbial (@ADVL), premodifing attribute (@NN>), or postmodifing attribute (@<NN).

During the last stage syntactic constraints are applied to words that remove the syntactic tags unsuitable for the context:

```
Aknas
   aken+s //_S_ com sg in **CLB // @ADVL
kustus
   kustu+s //_V_ main indic impf ps3 sg ps af #Intr //
@+FMV
tuli
   tuli+0 //_S_ com sg nom // @SUBJ
$.
```

Word form *aknas* was analyzed as an adverbial, the word form *tuli* was analyzed as the subject, and the verb *kustus* received the tag of a finite predicate.

In 1995 when we launched preparatory activi-

ties for the automatic syntactic parsing of Estonian, the Constraint Grammar was beyond doubt the most efficient grammar model for morphological disambiguation. The syntactic description of the CG was not as deep as in the case of other rule-based grammar models, but the CG output contained far less mistakes. CG has maintained that leading position from its introduction to the present day.

3 Method of Elaborating Rules

To elaborate the **morphological disambiguation constraints**, we established the more frequent groups of ambiguities. We found both more frequent ambiguous word forms as well as the ambiguous grammatical categories (the past participle with the interpretation of the adjective either in singular or plural, the noun or the verb; noun in the nominative, genitive and partitive case; noun in the genitive, partitive and aditive; adverb and adjective in the ablative case, etc.) (Puolakainen 01). This frequency table indicated which phenomena needed to be handled first of all. For each case samples were collected from text corpora and upon these observations tentative rules were compiled. The rules were tested against a manually tagged corpus where the correct interpretations provided by linguists had the corresponding tags attached to them. Depending on the results the rules were revised and tested again until their impact was satisfactory.

It was quite easy to compile the **syntactic mapping rules**. For the majority of rules it was necessary to transform the linguistic rules taken from the Grammar of the Estonian Language (Erelt *et al.* 93) into a suitable computational grammar form. Then the manually tagged text and the parser output were compared and most of the exceptions detected.

Compiling the **syntactic constraints** (which remove the added tags) was not that simple. The first rules were derived from the grammar books: all kinds of rules of agreement check the agreement of the predicate with the potential subject in person and number, and the agreement of attributes with their heads in the case. Also a number of rules checking the case of the object were based on grammars.

When the grammar books were exhausted the rest of the rules had to be generated on the basis of intuition: to examine a partly analyzed text and write new rules that would reduce the remaining ambiguities. The easiest to compile were rules establishing the complements of quantifiers and adpositions. For example, a word in the genitive case is a complement to a postposition if the postposition is immediately next it and it requires the genitive case.

Among the attribute rules, the simplest are those seeking whether in the left or right context there exists at all a word they may complement, as well as numerous rules checking the agreement or non-agreement. A number of rules are clearly of 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. While observing the word order in the sentence, it became obvious that such combinations as 1) object in the nominative or genitive case – predicate – subject in the nominative case, or 2) object in the nominative or genitive case – subject in the nominative case – predicate occur very rarely.

We tried to group the rules 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 in turn have been divided into groups based on their reliability.

The rules have been written in a way to reduce the amount of errors in the training corpus. The rule may go wrong only in exceptional cases or due to a mistake caused by another rule. While writing the rules we had in mind the objective of obtaining 99% recall of the parser on the training corpus.

4 Rules in the Constraint Grammar of Estonian

4.1 Rules of morphological disambiguation

There are 1,240 rules in the morphological disambiguation grammar half of which treat concrete word forms. The order of applying the rules is important. For example, it is essential to apply the phrasal verb rules before applying other rules as these determine most appropriately the part of speech of doubtful adverbs. As it is not possible to present the whole list of rules in this paper, we will bring some examples from different groups.

Phrasal verb rules

The gist of the problem lies in the fact that

many adverbs (affixed adverbs) in the phrasal verb may in certain context be either adpositions (*üle* – adverb in the phrase *üle elama* /to survive/ and preposition in the phrase *üle jõe* /across the river/), or coincide with some case of the noun (*kätte* – adverb in the phrase *kätte juhatama* /to indicate to/ and the noun *käsi* /a hand/ in the inessive case). This group is represented, for example, by the rule: select the adverb reading of the word form *kätte* if either to the left or to the right of this word in the clause occur forms of the verbs *võtma* or *juhatama*. E.g., *juhatas tee kätte* /indicated the way/ (adverb) but *pani kindad kätte* /put the gloves on/ (noun).

Determining the case of a declinable word based on the preceding numeral

The cardinal (except *üks* /one/) as well the word *mitu* /several/ in the nominative case is followed by a noun, pronoun or adjective in the partitive case, in other cases there has to be agreement between these words. For example, *kaks (_N_ card sg nom) maja (_S_ sg nom gen **part**)* /two houses/ and *kahe (_N_ card sg gen) maja (_S_ sg nom **gen** part)*.

Selecting between the noun and the adjective

This very frequent ambiguity is typically represented by words like *haige* /ill/, *vana* /old/; *kaotaja* /looser/, *võitja* /winner/; all the past participles.

Choose the part of speech noun if there is a preceding disambiguated adjective in the clause, there are no punctuation marks nor a verb allowing a predicate and no other nouns or pronouns in the clause, e.g. *rasket (_A_ sg part) haiget (_S_ sg part)* /gravely ill/.

The following is a selection of rules pertaining to certain word forms.

The word form *tuli* /fire, came, had to/ is provided by two interpretations by the morphological analyzer: verb form (*came*, *had to*) and noun (*fire*). The word form is very probably a verb if there is either a preceding or a following *da*-infinitive form in the clause, eg *tuli (_V_ main indic impf ps3 sg ps af) lugeda (_V_ main inf)* /one had to read/.

The word form *kui* /if, as, how/ may be either adverb (*how*) or conjunction (*if, as*) . In disambiguating this form we proceed mainly from the fact that the adverb *kui* must be followed by either an adverb or an adjective.

Choose the conjunction interpretation if *kui* is immediately preceded by a comma, the last but one word form is *siis* /then/ or if the right context contains the word form *siis*. For example: *Kui (Conj) homme sajab, siis jätan minemata* /If it rains tomorrow then I'll not go/, as well as *Siis, kui (Conj) homme sajab, jätan minemata*.

Choose the adverb interpretation if the following word is one of the list: *mitu* /a number of/, *vähe* /little, few/, *palju* /a lot of, much/, *hea* /good/, *halb* /bad/, *hästi* /well/, *kaua* /long/, *vana* /old/, *noor* /young/, *raske* /difficult/, *kerge* /easy/. For example, *Kui (adverb) kaua see kestab, ei tea keegi* /Nobody knows how long it will last/.

4.2 Clause boundary detection rules

The clause boundaries 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. In EstCG there are 47 clause boundary detection rules, a number of them for very specific cases.

4.3 Morphosyntactic mapping rules

27 syntactic tags of EstCG represent syntactic functions of traditional Estonian grammar (Erelt *et al.* 93), although there are some modifications considering the specialities of Constraint Grammar: CG annotates every word with some syntactic label while a 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 (complements of predicate). 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 verb chain is marked by five labels: finite or infinite, auxiliary or main verb, and a label for negation.

Syntactic tags are added to words by 180 morphosyntactic mapping rules. While compiling these it was taken into account that each rule should be optimal – there should be a correct tag among the tags added by the rule, and the rule should avoid adding rarely occurring tags to all words. This has been achieved by adding con-

text conditions to the rules. Thus, the adposition tags @P> or @<P will be added only if the clause contains an adposition requiring that specific case. After applying the mapping rules the percentage of syntactic ambiguity is typically very high (e.g., during our experiments, we observed that each word form had 3.8 tags on the average), and usually only verb finite forms and conjunctions remain unambiguous. At the same time, the error rate at this stage is very low, normally below 0.2%. The errors have been mostly caused by adjectives used in the role of nouns.

The verb rules have been separately compiled for auxiliary verbs, modal verbs, all types of supinum, gerundium, *vat*-infinitive, participles, and *da*-infinitive, and the rest of verbs are tagged as components of the finite predicate (@+FMV).

The noun rules are divided by cases, whereby it is checked whether the noun may belong to the complement phrase with specified government mode. The rules for pronouns and ordinals are very similar to those of nouns but contain certain exceptions.

The adjective rules mostly take into account whether the adjective may be a predicate complement or not. There are separate rules for adjectives performing the role of a noun, such as *vähene* /small/ or *viimane* /the last, latest/. The rules for ordinals are similar to those of adjectives. In addition to that, there are rules for adverbs, adpositions, conjunctions, abbreviations, and exclamations.

4.4 Syntactic constraints

EstCG has presently 1118 syntactic constraints.

Predicate grammar

The ambiguity of the predicate is caused by the past participles, *da*-infinitive, supinum, and *vat*-infinitive. The rules for solving these ambiguities can be divided into 3 groups:

1. Rules for checking the existence of auxiliary or modal verbs in the clause.
2. Rules for checking whether the verb causing the ambiguity is the sole verb in the sentence – in that case it is probably a predicate.
3. Rules based on the construction of a specific sentence which either remove or retain the predicate tag depending on the word order and lists of different verbs.

Subject grammar

Nouns, pronouns and numerals in the nominative or partitive case, and *da*-infinitives may occur as the subject. The adverbs functioning as quantifiers and adpositions as heads of an adposition phrase may perform the function of the subject.

The subject grammar does not solve all the subject ambiguities. The grammar rules for the object, predicate complement, attribute, and adverb quite often participate in determining the subject. They either select another function for the word tagged as the subject, or correspondingly remove the tags for the object, predicate complement, adverb, or attribute, leaving the single tag of the subject.

Object grammar

Nouns, pronouns and numerals in the nominative, genitive and partitive case, *da*-infinitives, *vat*-infinitives, and adverbs as heads of the quantifying phrases may occur as the object.

For determining the object more efficiently, we have employed lists of verbs. A list of intransitive verbs has been compiled based on the (Erelt *et al.* 93) and the corpus of Estonian, as well as lists of such verbs that allow the *da*-infinitive object or the object only in the partitive case.

Attribute grammar

It covers the major part of EstCG, more than half of the rules (587 rules). It is due to the fact that, on the one hand, the attribute tags are attached to almost all the word forms; on the other hand, for adjective and various pronouns one has to use 28 rules (14 cases, singular and plural) to check the agreement. There exist separate rules for determining adjectival and nominal pre- and postpositional attributes, pronoun, adverb, and adposition attributes; also, there are some rules for determining the infinite verb as the postpositional attribute.

Adverbial grammar

The composed rules pertain to nouns, infinitive verbs, and adjectives. As we have found, it is easier to detect adverbials by removing adverbial tags from non-adverbials, rather than to write rules that attempt to discover adverbials themselves. For example: the noun in the nominative case is not an adverbial if it a) is not included in the list of nouns denoting time (*aeg* /time/, *päev* /day/, *aasta* /year/, etc.); b) is not the numeral; c) is not a numerical quantifier; d) is not the sentence complement (e.g. the word *tõsi* /true/); e) is not immediately followed by the words *hiljem* /later/, *varem* /earlier/, or *kauem* /longer/; f) is not preceded by the pronoun *iga* /every/; g)

is not preceded by the ordinal; h) is not part of a comparison (there are no such words as *nagu* /as/, *kui* /as/, *otsekui* /as if/, *justkui* /as if/ to its left).

5 Results

5.1 The corpora used

To compile and test the EstCG, a 20,314 word corpus of Estonian was used that was manually tagged for morphological and syntactic features (hencetoforth referred to as the training corpus). The corpus consists of six original fiction texts, each from a different author (12,223 words altogether, derived from the Corpus of Written Estonian), a 6,373 word translation of a fiction text ("1984" by G. Orwell), and a 1,719 word newspaper text.

To assess the effectiveness of the parser, a manually tagged 9,663 word test corpus was used that had not been applied for optimizing nor evaluating the grammar rules.

5.2 Results of morphological disambiguation

The use of the disambiguator reduced the percentage of ambiguous word forms 4 times in both the training and the test corpus.

The results of morphological disambiguation in the training corpus are the following: precision is 83.39-89.68%, recall is 97.87-99.16%, and the percentage of words with one reading is 88.67-91.74%.

The results of morphological disambiguation in the test corpus are the following: precision is 85.49-89.16%, recall is 97.95-98.36%, and the percentage of words with one reading is 88.67-91.96%.

Most of the mistakes occurred when the cases of nouns, adjectives, and pronouns were determined: the difficulties occur in differentiating the nominative, genitive, partitive, and aditive cases. It was also the leading ambiguity type in the initial text and in remaining ambiguities after disambiguation as well. The second place in the errors toplist was taken by the problem of determining whether past participles were used as adjectives or verbs, and the third is closely related to the second one – determining whether the form of the verb *olema* /to be/ is the main or auxiliary verb.

5.3 Results of syntactic analyzer

The grammar of syntactic analyzer was tested on two types of corpora: the first one was manually morphologically disambiguated, i.e. all the errors of previous stages of analysis were fixed and morphological ambiguities removed by linguists. This type of corpus helped to determine the problems that are specific for syntactic disambiguation only. The second corpus consists of same texts but the preceding analysis is achieved automatically.

While compiling the grammar the goal was set to obtain as disambiguous analysis as possible on the training corpus, at the same time maintaining a high rate of recall (at least 98.5%). The recall of all texts did not fall below 98.9% in the training corpus, whereas the ratio of unambiguous analyses differs more than 7%, from 86.1% to 93.4% (precision is 89.61% for the whole corpus). It is lowest for the newspaper text which is stylistically a lot different from the others.

The results of syntactic disambiguation in the test corpus were a bit less promising but the difference was not too big: 0.7% reduction in recall and 2% reduction in precision. When analyzing the results of the test corpus it should be noted that the journalistic text is more difficult for the parser. The remaining ambiguity mostly consists of two tags, but single words may have from 3 to 5 syntactic tags. The largest class of ambiguities is formed by adverbials and adverbial attributes. 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. 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 appears 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.

Having analyzed the errors, we found that there were no systematic misinterpretations in the grammar. The errors are mostly caused by ellip-

sis, some errors occurred during determination of apposition and the third biggest group of errors exists in sentences where one clause divides the other into two parts. Most of the errors can be avoided by refining the contextual conditions of the rules. The rate of erroneous analyses is as well below 2% that was set as the goal in compiling the grammar.

The results obtained from the automatically morphologically disambiguated corpora are supposedly worse. In the test corpus, the recall of syntactic disambiguation was 95.50-96.85%, precision 76.36-79.22%, and the unambiguity 81.34-83.63%.

The analysis of errors on that type of corpus demonstrates that the majority of the additional errors have directly been caused by the errors of the morphological disambiguator and by words unrecognised by the morphological analyzer.

6 Conclusion

Constraint grammars have been written for the Basque, English, Norwegian, Portuguese, Swedish, Turkish, and now for the Estonian language.

The computational grammar and the parser elaborated during our project are the first attempts to automate the syntactic analysis of Estonian. The 2,500-odd rules of the Constraint Grammar of Estonian have been formed on the basis of the standard linguistic grammar of Estonian and the texts from the corpus of the written Estonian. The parser has been successfully used in two prototypes of practical applications: recognizing the noun phrases and the automatic generation of text summaries. The promising fields of application include information retrieval (taking into account the typical patterns of sentences while dealing with a specific topic, e.g., (Bangalore 97), text-speech generation (partial syntactic analysis for determining the intonation of the sentence), grammar checking in text editing, and translator aids (e.g., search for standard sentence constructions from a parallel corpus).

The directions of our future work are as follows.

- The share of the lexicon in the grammar must be increased. We need a lexicon containing certain semantic information covering, first of all, noun quantifiers and nouns that can fulfill the functions of adverbials of time and manner occurring in other cases than the nominative, genitive, and partitive. The problem of complex and phrasal verbs needs solving as well. At the moment their nominal and adverbial components are analyzed according to their morphological form, but it would be necessary to relate them to the verb. To do so, we need exhaustive lists of complex and phrasal verbs (this work has already been launched, cf. (Kaalep & Muischnek 02)).
- The volume of the tagged training corpus has to be increased.
- To experiment with different statistical methods for the automatic generation of grammar on the bigger training corpus.
- To expand the amount of syntactic tags. A more detailed tagging should considerably increase the efficiency of the object analysis.
- In the longer perspective we foresee a need for transition to a deeper description of the syntax, e.g., to take over the principles of the Functional Dependency Grammar (Järvinen & Tapanainen 97).

References

(Bangalore 97) Srinivas Bangalore. *Complexity of Lexical Descriptions and its Relevance to Partial Parsing.* Ph.d. diss., University of Pennsylvania, 1997.

(Erelt *et al.* 93) Mati Erelt, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, and Silvi Vare. *Eesti keele grammatika. II Süntaks.* Eesti TA Keele ja Kirjanduse Instituut, Tallinn, 1993.

(Hennoste *et al.* 98) Tiit Hennoste, Mare Koit, Tiit Roosmaa, and Madis Saluveer. Structure and usage of the Tartu University corpus of written Estonian. *International Journal of Corpus Linguistics*, 3(2):279–304, 1998.

(Järvinen & Tapanainen 97) Timo Järvinen and Pasi Tapanainen. A dependency parser for English. TR 1, Department of General Linguistics, University of Helsinki, 1997.

(Kaalep & Muischnek 02) Heiki-Jaan Kaalep and Kadri Muischnek. Using the text corpus to create a comprehensive list of phrasal verbs. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 101–105, Las Palmas, 2002.

(Kaalep 96) Heiki-Jaan Kaalep. Estmorf. a morphological analyzer for Estonian. In *Estonian in the Changing World*, pages 43–98. Tartu, 1996.

(Karlsson *et al.* 95) Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. *Constraint Grammar: a Language Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin and New York, 1995.

(Müürisep 00) Kaili Müürisep. *Eesti keele arvutigrammatika: süntaks.* Ph.d. diss., University of Tartu, 2000.

(Puolakainen 01) Tiina Puolakainen. *Eesti keele arvutigrammatika: morfoloogiline ühestamine.* Ph.d. diss., University of Tartu, 2001.