TREEBANK-BASED RESEARCH AND E-LEARNING OF ESTONIAN SYNTAX

Heli Uibo*, Eckhard Bick**
*University of Tartu (Estonia), **University of Southern Denmark

Abstract

The creation of syntactically annotated corpora of Estonian started at the end of 1990s with the training and test corpora for the Constraint Grammar shallow syntactic parser. By now the size of the Estonian Constraint Grammar Corpus is close to 300 000 running words. In 2004 the first attempts have been made to build deep syntactically annotated corpora (treebanks) for Estonian.

The Estonian treebanks have been annotated using the VISL formalism, developed at the University of Southern Denmark, which combines the phrase structure grammar with syntactic functions. There exist VISL treebanks for more than 20 languages already and they are used mainly for two purposes – for research and for education. Online visualization and query tools as well as edutainment software have been created within the VISL project to facilitate the both usages of treebanks.

Two small treebanks of Estonian were created as the joint work of University of Tartu and University of Southern Denmark. The research-oriented treebank Arborest has been semi-automatically derived from a section of the Estonian Constraint Grammar corpus and the Estonian teaching treebank was annotated manually. Currently, the Estonian teaching treebank consists of 100 sentences. The Estonian treebank Arborest contains 2500 sentences, 149 from which have been manually revised.

Keywords: annotated corpora, treebanks, syntax, Constraint Grammar, e-learning

1. Introduction

Data-driven methods are gaining more and more popularity and success in natural language processing. First, it is cheaper to let the computers discover the language rules instead of hand-crafting them, and, on the other hand, only the language software that contains probabilistic components, trained on the real data, could be able to cover the wide spectrum of texts produced by different language users and having different communicational goals.

For machine learning of syntactic rules of a particular language we need syntactically annotated corpora, also called *treebanks*. The term treebank comes from the graphical representation of a parsed sentence as a tree (cf. Figure 1). However, not only corpora containing tree-shaped representations are considered treebanks, but corpora with all kinds of structural analysis beyond the part-of-speech level, including semantic and discourse analysis (Nivre et al 2005). Treebanks can also be used for

training and evaluation of morphological and syntactical analyzers and human language technology applications, for e-learning and for linguistic surveys.

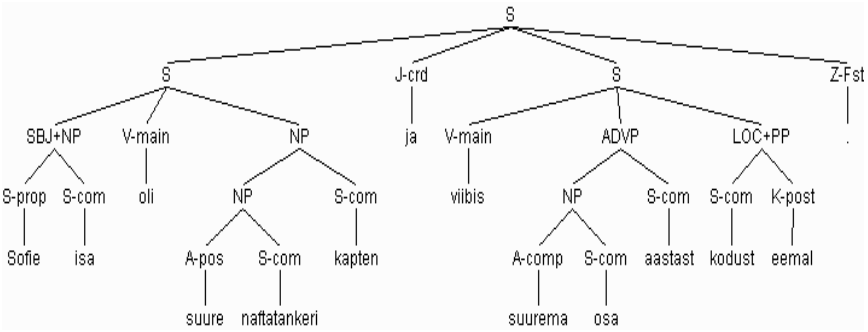

Figure 1. A sentence tree from the Estonian part of the Sophie Parallel Treebank

Creating a treebank is a time- and labour-consuming task. Therefore, it is important to look for the opportunities of re-using the existing methods and software for the treebank creation at its starting point. Ideally, the design of a treebank should be motivated by its intended usage, whether linguistic research or language technology development. However, in practice, there are a number of other factors that influence the design, such as the availability of data and analysis tools (Nivre et al 2005).

In 2003, before we started to create Estonian treebanks, we had a Constraint Grammar shallow syntactic parser for Estonian (Müürisep et al 2003) and some CG-annotated training and test material for the parser. In April 2003 the research group of computational syntax at University of Tartu joined the Nordic Treebank Network[1], a research network which aims at promoting treebank-related research in Nordic countries. The Estonian treebanks described in this paper have come into being thanks to the research cooperation between University of Tartu and other research institutions involved in this network. The most significant results have been gained thanks to the reuse of the method of semi-automatic creation of VISL treebanks developed at University of Southern Denmark (Bick 2003).

In this paper we will give an overview of treebank-related research activities for Estonian language. We will describe the creation process and annotation schemes of Estonian treebanks, which have been slightly different depending on the representation formats and purposes of the treebank. We will also look at the usage areas of syntactically annotated corpora of Estonian – research and e-learning.

2. Constraint Grammar Corpus of Estonian

The creation of syntactically annotated corpora of Estonian started at the end of 1990s. At that time we were developing a Constraint Grammar shallow syntactic parser for Estonian and some training and testing material was needed for that purpose. Depending on the funding the test corpus creation progressed in variable speed. By the time being the Estonian Constraint Grammar Corpus consists of ca 210 000 words of fiction, 80 000 words of newspaper texts and 6 000 words of legal texts. The corpus is

[1] http://w3.msi.vxu.se/~nivre/research/nt.html

stored as text files[2], each word on a separate line and its morphological and syntactical tags (Table 1) on the next line.

To use the existing treebank creation and usage tools it is often needed to make some kind of conversions on the corpus representation format. We have converted our syntactically annotated corpora to standard formats (NEGRA export format and TIGER XML) to facilitate the usage of treebank tools developed at University of Stuttgart within the TIGER project (Brants et al 2002).

Table 1. Estonian Constraint Grammar tag set

Syntactic function tag	**Meaning**
@SUBJ	subject
@OBJ	object
@ADVL	adverbial
@±FMV	finite (non-finite) main verb
@±FCV	finite (non-finite) modal or auxiliary verb
@AN>, @<AN	adjective as attribute
@NN>, @<NN	noun as a modifier (of a noun); apposition
@AD>, @<AD	adverb as a modifier (of a noun)
@Q>, @<Q	complement of quantor (*five men*)
@P> , @<P	complement of adposition (*on the table*)
@VN> , @<VN	participle as a modifier (of a noun)
@INF_N>, @<INF_N	infinitive as a modifier (of a noun)
@PN>, @<PN	adpositional phrase as a modifier (of a noun)

K. Kaljurand has written a Perl program for converting the CG corpus to NEGRA export format[3]. Now we can use the treebank creation tool Annotate and treebank search tool TIGERSearch to work on Estonian CG corpus. TIGERSearch is a powerful treebank search tool that imports grammatical structures from various formats, makes them searchable and displays them graphically. The graphical output of TIGERSearch is illustrated on Figure 2. Noticably, the CG annotated trees are very flat – phrase structure and the hierarchy of subclauses are not expressed.

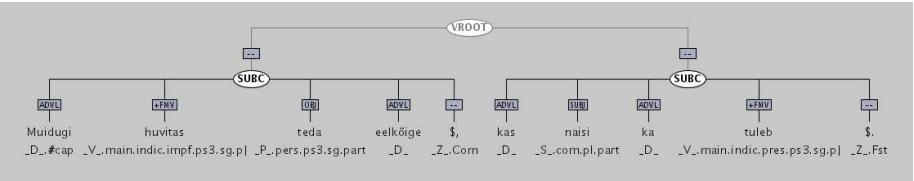

Figure 2. A sentence from the Estonian CG corpus in TIGERSearch graphical display format

There have been carried out some linguistically interesting experiments with Estonian CG corpus. Dependency relations were derived from the CG annotation, focusing on subject, object and adjective modification relations to build a database of

[2] http://lepo.it.da.ut.ee/~heli_u/SA/

[3] http://psych.ut.ee/~kaarel/Programs/Treebank/EstCG2Negra/

syntactically similarly behaving nouns. This database together with Estonian WordNet was used for word sense disambiguation (Kaljurand 2004).

3. Treebanks of Estonian

As the Constraint Grammar trees are too flat, we needed another formalism to represent the deeper structure of the sentence. The Estonian treebanks have been annotated using the VISL formalism, developed at the University of Southern Denmark, which combines the word-based shallow dependency tags with constituent trees (Bick 2003).

Each node in the VISL tree has two labels – a form label (e.g. fcl = finite clause, NP = noun phrase, VP = verb phrase, n = noun, adj = adjective etc.) and a function label (e.g. clause level syntactic functions like S = subject, O = object, P = predicate etc. and phrase-internal functions like H = head and D = dependent).

There exist VISL treebanks for more than 20 languages already and they are used mainly for two purposes – for research and for education. For both purposes small experimental treebanks of Estonian have been created.

4. Arborest – from shallow to deep syntax

The treebank Arborest was created semi-automatically. We have re-used the method of deriving phrase structure trees from the word-based Constraint Grammar annotation described in (Bick 2003). The Estonian treebank Arborest contains 2500 sentences, from which 149 have been manually revised. The evaluation of the automatically generated trees showed that 40 % of the trees were correct, i.e. had both correct branching structure and (almost) correct form and function labels.

Generalizing the study results, the correctly parsed sentence types were the following:

(1) simple sentences with the structures subject-predicate-object in any order plus optional adverbial(s) and predicative complement (C);
(2) composite sentences (subclauses bound with ja, ning, või, ehk or comma);
(3) complex sentences with subordinated clauses in the function of adverbial or object

Major sources of errors were adverbial attachment, non-finite subclauses, complex noun phrases and complex sentences with more than one subordinated clause.

More detailed description of creation and examination of Arborest is given in (Bick et al 2004). Estonian treebank Arborest together with a tgrep2-based search interface *tgrep-eye* is available at the webpage http://corp.hum.sdu.dk/arborest.html.

5. Online e-learning of Estonian syntax

VISL online edutainment games have been worked out at the University of Southern Denmark within the VISL project[4]. The games are based on the syntactically annotated corpora and are meant to learn the grammar of a particular language.

Using the VISL terminology some of the games are called "function games" (to learn syntactic functions) and others are "form games" (to learn word classes). The task to determine word classes (noun, verb, adjective etc.) has been implemented in different attractive ways in the games "Shooting gallery", „Labyrinth“„Wordfall“ a.o. In some other games, e.g. "Space rescue" the student has to determine the syntactic functions of words in a sentence.

[4] http://visl.sdu.dk

The teaching treebank of Estonian was annotated manually by H. Uibo, H. Nigol, K. Kerner and K. Nurmoja. It is a VISL treebank, as is Arborest, but we have decided to use a little different and in some aspects richer category set.

1. We have subcategorized the A (adverbial) tag:

- *fA* for clause level adjuncts
- *Ao* for NPs which are valency-bound to verb but are not objects as they are not in nominative, genitive or partitive case

2. As there exist both pre- and postpositions in Estonian, we invented a new tag *adp* for adpositional phrases (instead of *prp* used for tagging adp-s in the treebank Arborest).

3. Phrasal verb constructions are analyzed as a whole, using the tag *Vpart* as the function tag for the adverbial component.

Currently the Estonian teaching treebank consists of 100 sentences which are classified into the following ten complexity classes (10 sentences in each class):

1..10 S V and S V O (simple NPs, i.e. consisting from head only)
11..20 S V Adv and Adv V S (simple NPs)
21..30 like 1..20, but more complex NPs
31..40 S V O + adverbs allowed everywhere
41..50 like 1..40, but more complex VPs and predicatives (complements) allowed
51..60 simple sentences, having PPs in the adverbial function
61..70 composite sentences (subclauses on the same level in the tree)
71..80 complex sentences, subordinated clause in the function of object or adverbial
81..90 sentences containing coordinated NPs
91..100 sentences with non-finite subclauses

The teaching treebank and the VISL games for Estonian are available on the webpage http://beta.visl.sdu.dk/visl/et. A pilot project is carried on in some secondary schools to try out the e-learning of Estonian syntax in practice. We are going to revise the teaching treebank based on the feedback that we will get from the teachers and students. We are already aware of some difficulties of using the VISL games:

- User interface of the games is in English.
- Foreign words are used for word classes (in content, the material is suitable for 5.-7. year primary school students, but terminology is not known).

6. Conclusions and perspectives

By converting our corpora to the appropriate format we can reuse the tools developed for other languages. An important lesson learned in this process was that the use of standard encoding and annotation formats is crucial design decision from the very beginning.

In a longer perspective, we intend to extend the CG corpus to 500 000 running words, to correct and improve CG-to-PSG rules in the transformation grammar, and maybe refine the CG tagset (especially with regard to adverbial subclasses). A wider coverage of language variety is also desirable, and we would like to create spoken language treebanks for Estonian. How to represent dialogue act information etc is still open, but will hopefully be resolved in cooperation with the Nordic Treebank Network. On the teaching side, we intend to test the VISL games for Estonian in practice at some primary schools to get the feedback from students and teachers, and prepare the grounds for a more formal evaluation. Finally, our corpora are in continued need of manual revision, and a special focus area will be complexity classes of sentences, as well as adverbial subcategorization in both the Arborest and VISL teaching Treebanks. A

special corpus initiative is the Estonian part of the Sophie parallel Treebank, which is being created using the methodology described for Arborest.

References

Bick, Eckhard 2003. A CG & PSG hybrid approach to automatic corpus annotation, In: Simow, K.; Osenova, P. (eds.) *Proceedings of SProLaC2003* (at Corpus Linguistics 2003) Lancaster. 1–12.

Bick, Eckhard; Uibo, Heli; Müürisep, Kaili 2004. Arborest – a VISL-style treebank derived from an Estonian Constraint Grammar corpus. In: Kübler, S.; Nivre, J.; Hinrichs, E.; Wunsch, H. (eds.) *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004).* Tübingen. 1–14.

Brants, Sabine; Dipper, Stefanie; Hansen, Silvia; Lezius, Wolfgang; Smith, George 2002. The TIGER treebank. In: *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT 2002)*. Sozopol, Bulgaria. 24–42.

Kaljurand, Kaarel 2004. Word Sense Disambiguation of Estonian with syntactic dependency relations and WordNet. In: *Proceedings of ESSLLI-2004*, Nancy, France. 128–137.

Müürisep, Kaili; Puolakainen, Tiina; Muischnek, Kadri; Koit, Mare; Roosmaa, Tiit; Uibo, Heli 2003. A New Language for Constraint Grammar: Estonian. In: *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP 2003)*. Borovets, Bulgaria. 304–310.

Nivre, Joakim; de Smedt, Koenraad; Volk, Martin 2005. Treebanking in Northern Europe: A White Paper. In: *Nordisk Sprogteknologi. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004*. Copenhagen: Museum Tusculanums Forlag. (forthcoming)

HELI UIBO is lecturer of language technology at the Institute of Computer Science, University of Tartu and member of the informal research group of Computational Linguistics, University of Tartu. She received her M.Sc. (computer science) in 1999 at the University of Tartu. Her research interests include Natural Language Processing, especially morphological and syntactical analysis of Estonian and creation and usage of syntactically annotated corpora (treebanks). She is the site coordinator in the Nordic Treebank Network (2003-2005), a research network funded by NorFA. E-mail: heli.uibo@ut.ee.

ECKHARD BICK works as research lector at the Institute of Language and Communication, University of Southern Denmark, where he is project leader of the VISL project. Eckhard Bick has degrees in Medicine (University of Bonn, 1984), Nordic Languages and Portuguese (cand.mag., Århus University, 1993). He defended a dr.phil.-thesis in Linguistics in 2000, also at Århus University, on Constraint Grammar based automatic analysis of Portuguese, and has since written Constraint Grammars for a number of languages. Current interests include corpus annotation, treebanks, grammar based spellchecking, computer assisted language learning, named entity recognition and question answering. E-mail: eckhard.bick@mail.dk.