Syntactically annotated corpora of Estonian
Heli Uibo
University of Tartu
Tartu, Estonia
Heli.Uibo@ut.ee

Abstract
Syntactically annotated corpora are needed 1) to train and test parsers and various language technological products - grammar checkers, information retrievers and extractors, machine translators etc; 2) to check the agreement of existing linguistic theories with the real language usage. The corpora can be annotated on different levels of depth. In shallow syntactically annotated corpora a syntactic function is determined for every wordform; in deep syntactically annotated corpora (treebanks) also the dependency structure is determined for every sentence (graphically represented as a tree). There exists a Constraint Grammar shallow syntactic parser for Estonian, developed by K. Müürisep and T. Puolakainen. To train the parser, we have annotated texts of written Estonian (20 000 words of fiction, 6 000 words of legal text and 10 000 words of newspaper texts). By now we have extended the size of the corpus up to 200 000 words. We have also started to build two versions of Estonian treebank. A parallel corpus of 50 sentences from J.Gaarder's novel "Sophie's world" has been annotated and aligned and a Constraint Grammar plus phrase structure hybrid treebank is being developed, currently consisting of 2400 automatically generated trees, 149 of them manually revised.

1. Introduction

Linguistically annotated corpora are valuable, but quite labor-consuming type of language resource, unless we want to give up in its reliability. They are used for creation, training and evaluation of parsers and various language technological software tools. Language corpora can contain linguistic information of different levels of depth – on word level (full morphological analysis or Part-of-Speech tags), on sentence level (shallow or deep syntactic annotation), on the level of meaning (semantic roles and relations) or on the text level (discourse information). Usually the deeper levels of annotation presume the more shallow ones.

The richer the linguistic annotation is, the wider is the range of possible usages of the corpus. For instance, if morphologically tagged corpus is usable for the development of morphological analyzers and spelling checkers, then morphologically and syntactically annotated corpus can be applied as a test material both for spelling and grammar checkers, automatic summarizers, information retrieval and extraction tools. For development of a machine translation system, furthermore, morphologically, syntactically, semantically and pragmatically annotated parallel corpora would be ideal.

Among the syntactically annotated corpora the difference is made between shallow and deep syntactically annotated corpora, or treebanks. A treebank is a syntactically (sometimes also semantically) annotated text corpus, where the structure of sentences is given, so that it can be represented as a tree (directed acyclic graph with a single root node).

The trees can be either phrase structure or dependency trees. The kinds of syntactic information included into treebanks may vary: sometimes only syntactic functions (like subject, predicate, object, adverbial, etc.) are given, sometimes – the constituent structure of a sentence, or phrases (noun phrase or NP, verb phrase or VP, adverbial phrase or AdvP, prepositional phrase or PP, etc.), and often both, e.g. phrase names are the node labels and syntactic functions are the edge labels of a tree. Sometimes the generalization is made from treebanks to "graphbanks", e.g. if secondary edges are used to represent ambiguous parses. It is up to the treebank builders and of course it depends on the purpose of the treebank what kind of information to include. Usual conventions also include the eligibility of discontinuous structures and ambiguous analyses.

To create, maintain and use syntactically annotated corpora, various tools are needed.
The creation tools include parsers and text-based or graphical (more intuitive!) tree editing tools. Examples of usage tools are parser evaluation programs, corpus search and visualization tools. The importance of corpus conversion tools is also hard to overestimate, as there is no standard for the representation format yet and conversions are needed to unify treebanks or to use the tools developed for another corpus representation format. By the time being, XML has probably been the most widely used language for representation of trees.

In this paper we will give a brief overview of the stage of development of syntactically annotated corpora for Estonian language. At the moment there exist a shallow syntactically annotated corpus of considerable size and two small-scale experimental treebanks for Estonian.

2. The Constraint Grammar Shallow Syntactically Annotated Corpus

The shallow syntactically annotated corpus was considered necessary for training and evaluation of the Constraint Grammar based shallow syntactic parser of Estonian, written by Kaili Müürisep (Müürisep 2000). The development of the corpus started in 1998 with the gold standard corpus, consisting of 20 000 words of Estonian original fiction from 1980s. During 1999-2003 the corpus has been extended to ca 200 000 words, including 177 000 words of fiction, 10 000 words of newspaper texts and 6 000 words of legal texts.

The process of syntactic annotation was as follows:

1) The excerpts of ca 2000 words were chosen from the Corpus of Written Estonian (http://www.cl.ut.ee) and morphologically tagged, using the tagger by H.-J. Kaalep.

2) Morphologically analyzed texts were parsed with Constraint Grammar shallow syntactic parser CGP (Constraint Grammar Parser), which includes rules for both morphological disambiguation (Puolakainen 2001) and determination of syntactic functions (Müürisep 2000). Currently the precision of the syntactic parser is 76.4-79.2 % and recall 95.5-96.9 %. Because the precision and recall can never gain 100%, the annotation has to be manually revised and disambiguated.

3) Each automatically parsed text was revised by two linguists in parallel to avoid tendencies drawn by subjective language cognition. A human has to solve the cases, where the parser
   a. has determined several syntactic functions for a word form,
   b. has not determined a syntactic function,
   c. has determined a wrong syntactic function.

   There has been developed a GUI based annotator's tool to make the annotator's job more convenient.

4) The results were compared automatically by Unix scripts.

5) The cases where annotators disagreed were discussed.

6) The final version of the annotated text was created.

The most difficult cases which caused a lot of disagreement between annotators have been the choice between adverbial and noun phrase modifier, the analysis of quantifier phrases, determination of subject and predicative in the sentences of type [Eesti pealinn] (SUBJ or PRD?) on [Tallinn](PRD or SUBJ?) (Tallinn is the capital of Estonia), distinction between subject and object (synchronously participle and predicative), and to determine function for infinitive verb forms (Muischnek et al 2000).

There exist a webpage for the project of shallow syntactically annotated corpus http://math.ut.ee/~heli_u/syntkorpus.html. There is a directory of annotated texts available on the site as well as pointers to the guidelines for annotators and tools and interfaces connected to the corpus – GUI-based annotator's tool by K. Pirn and syntactic dictionary by K. Kaljurand. The work in progress includes the extension of the corpus as well as consistency check in the existing corpus.

3. Sofie Parallel Treebank

The Sofie Parallel Treebank is being developed as a test-bed for pilot studies on treebanks inside the Nordic Treebank Network1, the network funded by NorFA language technology program and joining 15 academic institutions from Sweden, Norway, Denmark, Finland, Estonia and Iceland to

1 http://www.masda.vxu.se/~nivre/research/nt.html
promote the treebanks-related research in Nordic countries.

The Sofie Parallel Treebank will include the first chapter of Jostein Gaarder's novel "Sophie's World" in all the languages contributing to the network. Currently, the parallel treebank includes Swedish, German, Norwegian, two versions of Danish and Estonian, 50-100 first sentences annotated from each language.

The syntactic structure represented in the trees of different languages is not similar – there exist Discontinuous Grammar dependency treebank and VISL\textsuperscript{2}-style phrase structure treebank for Danish, dependency treebank for Swedish, NEGRA-style treebank for German, phrase structure treebank for Norwegian and Penn-style phrase structure Treebank for Estonian in the Sofie Parallel Treebank. Only the representation format of trees is uniform – TIGER XML.

For creation of Estonian Sofie Treebank the tool Annotate\textsuperscript{3} was used. Annotate is a tool for semi-automatic annotation of treebanks, developed inside the NEGRA project. Terminal nodes, non-terminal nodes, and edges of a tree can be labeled. For Estonian Treebank the Penn Treebank tag set was used. Terminal nodes were read in as previously tagged with parts-of-speech tags, non-terminal nodes were labeled with phrase categories as NP, VP, PP, and edges were labeled with grammatical and semantic functions like SBJ (subject), PRD (predicative), LOC(location), TMP(temporal), etc. Annotated corpus was first saved in NEGRA format in a relational database and then converted to TIGER XML format.

We have not been satisfied with the Penn II tag set, because two very important function labels were missing – object and the head of phrase. Probably, we will rebuild the Estonian Sofie Treebank, using the method, which we will describe in the next section.

The Sofie Parallel Treebank can be examined and tgrep-searched at the password protected website http://omilia.uio.no/sofie.

4. Arborest – Experimental VISL-style Hybrid Treebank for Estonian

There have been carried out experiments on creating an Estonian treebank semi-automatically in cooperation with dr. Eckhard Bick from the University of Southern Denmark. He is the leader of VISL project, aimed in developing interactive language learning tools for 22 different languages and including smaller or larger treebanks (Bick 2003) for every language as well.

The trees were generated automatically from a sample of the Constraint Grammar shallow syntactically annotated (parsed and manually corrected) Estonian text corpus, described in section 2. There has been written an experimental phrase structure grammar module by E. Bick with help from H. Uibo and K. Muischnek, generating VISL-style phrase structure from CG-annotated text.

149 trees were examined visually, using the Slanttree Treebank viewer\textsuperscript{4}, developed inside the VISL project. It turned out that the CG-to-PSG rules generated 50 absolutely correct trees (i. e. having both correct structure and correct labels for forms as well as for functions). Probably the ratio of correct structures is even higher, as there has been occurred a systematic syntax error caused by the format of representation of derived word forms in morphological analysis in ca 30 sentences. After manual revision of the present set of trees the changes will be made to the PSG rules and the updated rules will be applied to another sample from the CG-annotated corpus.

The experimental VISL-style Estonian Treebank can be searched at the webpage http://corp.hum.sdu.dk/arborest.html.

5. Conclusion and Perspectives

There exist three syntactically annotated corpora for Estonian – 200 000 words of Constraint Grammar shallow syntactically annotated texts, a Penn-style phrase structure treebank of 50 sentences and a form-function VISL-style hybrid treebank of 2500 sentences (first 149 sentences manually revised).

The work in progress includes the enlargement of all the corpora. The research is particularly concentrating on improving the CG-to-PSG rules to facilitate the easy semi-

---

\textsuperscript{2} http://beta.visl.sdu.dk/visl2/

\textsuperscript{3} http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

\textsuperscript{4} http://beta.visl.sdu.dk/visl/Slanttree-doc.html
automatic way of building an Estonian treebank. We are also planning to create a syntactic-semantic dependency treebank for Estonian, which will be semi-automatically generated, using the algorithm given in (Nivre 2003) from one of the existing experimental phrase structure treebanks.

6. Bibliographical References


