

Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies

Kadri Muischnek, Kaili Müürisep,
Tiina Puolakainen

Central topic

Dependency treebank in Universal Dependencies formalism adapted from an existing dependency treebank for Estonian. this adaptation was doen semi-automatically using a Constraint Grammar transfer rule system.

Methodology

Structure of annotations

The Estonian Dependency Treebank (DT) is annotated in Constraint Grammar style. There are three layers:

- morphological

- surface syntactic
- dependency

This is an example word tag in a larger sentence (for more information see Figure 1 in the paper).

```
"<lamnast>"
```

```
"lamnas" Lt S com sg part @<Q #6->5
```

“ The used set of syntactic relations derives from Constraint Grammar, but the definitions of syntactic relations...are based on an academic description of Estonian grammar

Differences between UD and EDT annotation

Both EDT and UD adopt dependency grammar-based annotation guidelines. However, different syntactic relations are used and some phenomena are analyzed differently.

POS tags

No DET tag in estonian UDT, smilar decision made for The Fininnish UDT. PART not used because these things are currently tagged as adverbs or pronouns and it would require manual effort to retag them.

No discussion of annotation of morphological features.

Ditransitives are not used as there are no grammatical descriptions of Estonian that describes ditransitives in Estonian.

EDT distinguishes between finite and non-finite (*subordinate*) clauses with finite clauses not indicating the syntactic relation between the head of the finite clause and the main clause *what are they doing here then? This is very unclear in this paper, maybe I need to read the paper for the EDT in order to make sense of this.*

EDT annotated modals and other auxiliaries as multi-word predicates. Many of these are set up as complementary clauses with ccomp and xcomp in UD instead.

Primacy of content words in UD causes a large number of changes. EDT did a lot of relations between functional words. For example, nouns in a prepositional phrase were dependents of the preposition, while the preposition was dependent on the larger context. In UD, this has to be changed because dependency relations need to be between content words.

Conversion procedure

- Rearrange subtrees, find connections between UD and EDT *I thought this was manual exploration of differences first, but it does appear this is the actual tree rewriting*
 - Using Vislcg3 *like I intend to*
- Convert from CG3 format (default in ED%) to CONLL-U, convert pos tags, morphological features using simple mapping.
- Formal checks to verify there is one and only one root, verify valid dat, all fields filled in.

Findings

Estimation of conversion quality:

- Used MaltEval
- UAS of 96.3
- LAS of 98.4
- annotation of punctuation marks was an issue.
- ccomp is the most error prone dependency relation at 64.3%

UD's emphasis on dependencies between content words results in projectivity (often). Where EDT was non-projective, the UD version is projective.

Follow up readings

Revision #1

Created Fri, Apr 24, 2020 4:35 PM by [kenneth](#)

Updated Fri, Apr 24, 2020 4:41 PM by [kenneth](#)