Project proposal

The inter-clause dependencies in the Estonian UD Treebank

Eleri Aedmaa

1 Purpose and aims

As a part of the Universal Dependencies (UD) project\(^1\) the first version of the Estonian UD Treebank\(^2\) has created by Jan Štěpánek and Daniel Zeman. The current version is taken from the HamleDT 3.0\(^3\), which is based on the VISL-style hybrid Estonian Treebank, also called Arborest, created by Bick et al. (2004). This relatively small dataset will be replaced by the bigger and latest treebank called Estonian Dependency Treebank (EDT) (Muischnek et al., 2014). The conversion of the EDT to the UD is work in progress for the authors of the EDT. The goal of this project is to contribute to the conversion of the EDT to the UD through the development of the inter-clause dependencies for the Estonian UD Treebank. The result of this project will be evaluated as a part of the general evaluation of the conversion of the EDT to the UD.

2 Survey of the field

As mentioned previously, the current UD treebank for Estonian is taken from HamleDT (Zeman et al., 2014) and is originally based on Constraint Grammar (CG) and Phrase Structure Grammar (PSG) (Bick et al., 2004). The experiments showed that the PSG suits well for representing the Estonian noun phrase, but as the components of a multi-word verb may be separated from each other by intervening constituents, Estonian has no proper verb phrase (Muischnek et al., 2014). It is also claimed that the dependency representation is more suitable for languages with variable word order (Nivre, 2005), so the decision to build the corpus for syntactic dependencies was made. The EDT is now a 400,000-word corpus annotated for dependency syntactic structures. The annotation scheme for the EDT is somewhat inspired by the Stanford typed dependencies representation (SD) (de Marneffe and Manning, 2008), but a lot of information that the SD tagset presents explicitly in the form of syntactic labels is presented in the EDT as a combination of morphological and syntactic labels. The EDT morphological annotation scheme is slightly different from Universal Dependencies Scheme (UDS): the EDT lacks POS tag for determiners (substituted by pronouns) and also some universal features like gender, animacy, aspect, definiteness or state. Furthermore, while annotating the dependency relations that hold between the clauses, it is only stated that there is a dependency relation between the clauses, without any labels marked (Muischnek et al., 2014). Besides, in the HamleDT the dependencies in Arborest are labelled, but for the EDT it needs to be done, which makes the conversion of the EDT to UD more difficult.

The first release of the UD contained ten languages; one of them was Finnish, which belongs to the Finno-Ugric language family and is closely related to Estonian. Pyysalo et al. (2015)

\(^1\)https://universaldependencies.github.io/docs/
\(^2\)https://universaldependencies.github.io/docs/et/overview/introduction.html
\(^3\)http://ufal.mff.cuni.cz/hamledt
describe the conversion of the Turku Dependency Treebank (TDT) to UD. They introduced the mapping of the TDT annotation to the UD standard, described the challenges and their resolution, and evaluated a state-of-the-art parser on the TDT and on the UD Finnish treebank. They found that the performance of the parser improved in the conversion, which supports both the accuracy of the conversion and the feasibility of UD as a parsing target.

The difference between the conversions of the EDT and the TDT to UD lies on the fact that the annotation scheme of the TDT is based on SD (Haverinen et al., 2014), but the annotation scheme of the EDT is claimed to be just comparable to the SD (Muischnek et al., 2014). Hence, the EDT conversion to the UD should start with the comparison of the EDT and the SD annotation schemes as Bosco et al. (2013) did when they converted an existing dependency-based Italian treebank (MIDT) into the SD annotation formalism. They specialized the SD annotation scheme to deal with the peculiarities of the Italian language and built the new and very valuable Italian Stanford Dependency Treebank (ISDT).

In addition to referred studies, much work with UD for different languages has been done (e.g. for Persian (Seraji et al., 2014), Swedish (Nivre, 2014), Russian (Lipenkova and Souček, 2014), etc.). Based on the results by McDonald et al. (2011; 2013), it can be assumed that the development of inter-clause dependencies for the EDT could benefit not only from closely related Finnish UD Treebank, but also from the UD Treebanks for other languages.

3 Programme description

The development of the inter-clause dependencies for the Estonian UD Treebank will be based on the guidelines of the UD, that builds on Interset (Zeman, 2008), Google Universal Part-of-Speech Tags (Petrov et al., 2011), HamleDT (Zeman et al., 2012), Universal Dependency Treebanks (McDonald et al., 2013), and Universal Stanford Dependencies (USD) (de Marneffe et al., 2006; Tsarfaty, 2013; de Marneffe et al., 2014). The development of Estonian UD Treebank benefits greatly from the work by de Marneffe et al. (2014) because they proposed an improved taxonomy to capture grammatical relations across languages, including morphologically rich languages.

The first stage of the project contains further reading of related work, the study of the UD guidelines, and the observation of the existing Estonian UD Treebank in order to see how the inter-clause dependencies have been annotated there. Based on the findings it will be decided whether to correct the current UD Treebank for Estonian, or to convert inter-clause dependencies in EDT during the second stage of the project. The second part of the project includes the development of the dependencies (and the evaluation if possible).

The main purpose of this project is not to evaluate results, but the evaluation will be definitely a part of the general evaluation of the UD for Estonian. The evaluation could be carried out in the similar way to Pyysalo et al. (2015). As the EDT has been used for improving the existing rule-based parser, training MaltParser, and experimenting with the combinations of these two (Muischnek et al., 2014), there are parsers for which the results of the conversion of the EDT to UD could be roughly compared to.

4 Significance

This project is a part of the conversion of the EDT to the UD. The overall goals of this project are validating and extending the UD treebanks for Estonian and to contribute for the UD project in general. In addition, the UD Treebank for Estonian is important resource for the Estonian natural language processing tools as well as linguistic research.
References


