CG Roots of UD Treebank of Estonian Web Language

Kadri Muischnek
University of Tartu
Estonia
kadri.muischnek@ut.ee

Kaili Müürisep
University of Tartu
Estonia
kaili.muurisep@ut.ee

Dage Särg
University Of Tartu
Estonia
dage.sarg@ut.ee

Abstract

This paper describes a method building UD Treebank of Estonian Web Language from scratch. First, the texts were parsed using Estonian CG parser and the parser output was manually checked by two human annotators. After that, the CG annotations were converted into UD annotations by means of CG rules and external scripts. Apart from providing a detailed overview of this method, the paper also discusses benefits and limitations of this approach.

1 Introduction

This contribution reports on a project of building a preliminary version of the UD Treebank of Estonian Web Language (EWTB) and lessons learnt in the course of this effort.

Universal Dependencies (UD) is an open community effort to create cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework (Nivre et al., 2016).

As the Estonian UD Treebank (EDTB) has been part of the UD treebank collection since its Version 1.2 (Muischnek et al., 2014b), the corpus of web language has been included since Version 2.4. The main Estonian UD Treebank contains 30,723 trees, 434,245 tokens. EDTBs texts represent the "classical" genres of written language: fiction, newspaper and scientific texts. EWTB (1660 trees, 27,000 tokens) includes a small sample of texts from the corpus Estonian Web 2013.

The main aim of the UD effort is to facilitate developing better parsing techniques and better parsers. By "better" one also bears in mind better coverage of texts that are "out there" and need to be parsed for practical purposes. These texts include also the user-generated internet content containing a large variety of genres, differing from the normed language usage of the "classical" texts and also from each other in orthography, lexicon and even in the preferred syntactic structures. So we are extending the coverage of Estonian UD and as a pilot project we have annotated a small collection of web texts and published it as a UD Treebank of Estonian Web Language (EWTB) in UD Version 2.4.

In the UD repository different internet genres (blogs, web, social, reviews) are distinguished. Of those, EWTB contains blogs, social (forum posts) and other web texts, but no reviews.

2 UD Treebank of Estonian Web Language (EWTB)

EWTB includes a small sample of texts from the corpus Estonian Web 2013[1]. Estonian Web 2013 belongs to the so-called Ten-Ten corpus family. The texts have been crawled from the web, cleaned from non-textual material, tokenized and analysed morphologically (lemmatized). The same tools were used for tokenizing and lemmatizing classical written texts and more informal web texts, so the quality of the original morphological analysis was not reliable. Thus we preserved the tokenization but created new morphological annotation, including lemmas.

The creation of EWTB proceeded in two steps. First, the texts were annotated using the Estonian Constraint Grammar annotation scheme for morphological analysis and dependency parsing (Muischnek et al., 2014a). The annotation standard was the same as used for annotating the Estonian Dependency Treebank (Muischnek et al., 2014b), but one additional syntactic label has been introduced, namely that of discourse particle. The initial annotations were created using the Constraint Grammar parser for Estonian and the parser output was manually checked by two human anno-

[1] DOI:10.15155/1-00-0000-0000-0000-0011FL

tators. The preliminary Constraint Grammar style treebank of web texts is described by Särg et al. (2018).

The CG annotations were converted into UD annotations by means of Constraint Grammar rules. The conversion rules and conversion process are discussed in detail in Muischnek et al. (2016). Resulting UD annotations were again manually checked, but this time by one person. Also, several consistency checks were made using the Udapi tool (Popel et al., 2017).

Such a procedure - creating UD treebank by converting Constraint Grammar annotations into UD annotations - has also been used while creating the North Sámi UD treebank (Sheyanova and Tyers, 2017).

The annotation scheme of UD has been enhanced on each release, as well as the developers of the corpora are becoming more and more demanding for the correctness and consistency of the annotation.

When converting the new corpus from CG to UD, it appeared that, in addition to known problems in determining the function of clauses, it was necessary to check:

- POS tags of pronouns and their classification. Estonian CG annotation employs only pronoun part-of-speech, UD uses also determiners. Although most of the cases can be disambiguate rule-based, there are some cases which only human could solve.

- The annotation of names and appositions is different in CG and UD. The leftmost part of a multi-word name is the head in UD while Estonian CG annotates the last part of a multi-word name as the head. As for appositions, Estonian CG annotation scheme treats them as attributes.

- UD version 1 employed the same tag for nominal modifier of nouns and verbs, newer versions make the distinction. Fortunately, the conversion from CG is straightforward.

- The annotation of copula clauses is different in Estonian CG. Also, the definition of copula clause is wider as in CG and the straightforward rule-based conversion is not possible (Muischnek and Müürisep, 2017).

- The annotation of clausal complements (ccomp) and open clausal complements (xcomp).

- The annotation of elliptical constructions: rule-based detector can recognize some elliptical clauses but not all. Atypical elliptical clauses are quite frequent in the corpus of web language. Also, empty nodes have been included into UD syntax trees and the whole clause has an extra annotation of enhanced dependencies.

Figures 1 and 2 in Appendix 1 illustrate the format of CG and UD annotation of the EWTB sentence (1). The final version of the paper will discuss the transfer procedure in more detail.

(1) *Zopp ei löönud ühtki ässa , kuid ka*
Zopp not hit none ace , but too
vastane vaid ühe .
opponent only one .

Zopp did not hit any ace and his opponent hit only one.

3 Future plans and conclusion

The conversion rule set consists of approximately 1000 rules which transfer texts from CG format to UD. Some conversion steps need human knowledge and their rule-based automation is impossible (or hard). As for future research, we plan to increase the treebank and improve it by adding coreference annotation.

Acknowledgments

This study was supported by the Estonian Ministry of Education and Research (IUT20-56), and by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies).

References

Kadri Muischnek and Kaili Müürisep. 2017. Estonian copular and existential constructions as an UD annotation problem. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*, pages 79–85. Linköping University Electronic Press.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2014a. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. In *Baltic HLT*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 111–118. IOS Press.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian Dependency Treebank: from Constraint Grammar Tagset to Universal Dependencies. In *Proc. of LREC 2016*.

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014b. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291. University of Tübingen.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*. European Language Resources Association (ELRA).

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for universal dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Linköping University Electronic Press.

Dage Särg, Kadri Muischnek, and Kaili Müürisep. 2018. Annotated Clause Boundaries' Influence on Parsing Results. In *Proceedings: 21st International Conference on Text, Speech and Dialogue*.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pages 66–75.

Appendix

```
"<s>"

"<Zopp>"
        "Zopp" L0 S prop sg nom @SUBJ #1->3
"<ei>"
        "ei" L0 V aux neg @NEG #2->3
"<löönud>"
        "löö" Lnud V main indic impf ps neg @FMV #3->0
"<ühtki>"
        "üks" Lki P dem indef sg part @NN> #4->5
"<ässa>"
        "äss" L0 S com sg part @OBJ #5->3
"<,>"
        "," Z Com #6->6
"<kuid>"
        "kuid" L0 J crd @J #7->9
"<ka>"
        "ka" L0 D @ADVL #8->9
"<vastane>"
        "vastane" L0 S com sg nom @SUBJ #9->3
"<vaid>"
        "vaid" L0 D @DN> #10->11
"<ühe>"
        "üks" L0 N card sg gen @OBJ #11->9
"<.>"
        "." Z Fst #12->12
"</s>"
```

Figure 1: CG annotation of the sentence.

```
# sent_id = ewtb1_010703_23
# text = Zopp ei löönud ühtki ässa, kuid ka vastane vaid ühe.
1    Zopp   Zopp   PROPN  S    Case=Nom|Number=Sing   3    nsubj  3:nsubj _
2    ei     ei     AUX    V    Polarity=Neg  3      aux  3:aux  _
3    löönud lööma  VERB   V    Connegative=Yes|Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act   0    root   0:root _
4    ühtki  üks    DET    P    Case=Par|Number=Sing|PronType=Ind    5    det    5:det  _
5    ässa   äss    NOUN   S    Case=Par|Number=Sing   3    obj    3:obj  SpaceAfter=No
6    ,      ,      PUNCT  Z    _    9    punct  9.1:punct    _
7    kuid   kuid   CCONJ  J    _    9    cc     9.1:cc _
8    ka     ka     ADV    D    _    9    advmod 9:advmod     _
9    vastane vastane NOUN  S    Case=Nom|Number=Sing   3    conj   9.1:nsubj    _
9.1  lõi    lööma  VERB   V    Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act _    _    3:conj _
10   vaid   vaid   ADV    D    _    11   advmod 11:advmod    _
11   ühe    üks    NUM    N    Case=Gen|Number=Sing|NumType=Card    9    orphan:obj    9.1:obj _
12   .      .      PUNCT  Z    _    3    punct  3:punct _
```

Figure 2: UD annotation of the sentence.