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) and is freely available[2].

3 The conversion procedure

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.

3.1 Clausal dependencies

Estonian CG annotation scheme is quite fine-grained for annotating intra-clausal phenomena, and thus the transfer of annotation is not very complicated inside the clause. But although we annotate the dependency relations that hold between the clauses, our scheme does not distinguish the names of those relations, and the annotation only shows that there is a dependency relation between the clauses. That should be considered one of the main shortcomings of our CG annotation scheme. The heads of subclauses have been annotated by label *dep* and then corrected manually. Figure 1 illustrates the sentence (1) in the CG scheme and its correct syntactic counterparts in the UD schema are presented in Figure 2 (the column of morphological features is omitted). The label of the 5th token has been corrected manually.

(1) *Usutakse* *et* *puud* *suudavad*
believe-IMPRS that trees can
talletada *piisavalt* *CO*$_2$.
strore enough CO_2 .

It is believed that trees can store enough CO_2.

```
"<Usutakse>"
   "usku" Ltakse V main indic pres imps af @FMV #1->0
"<,>"
   "," Z Com CLB #2->2
"<et>"
   "et" L0 J sub @J #3->5
"<puud>"
   "puu" Ld S com pl nom @SUBJ #4->5
"<suudavad>"
   "suut" Lvad V main indic pres ps3 pl ps af @FMV #5->1
"<talletada>"
   "talleta" Lda V main inf @OBJ #6->5
"<piisavalt>"
   "piisavalt" L0 D @ADVL #7->6
"<CO2>"
   "CO2" L0 Y nominal ? @OBJ #8->6
```

Figure 1: CG annotation of the sentence.

1 Usutakse	usku	VERB	V	...	0	root	_	_
2 ,	,	PUNCT	Z	_	5	punct	_	_
3 et	et	SCONJ	J	_	5	mark	_	_
4 puud	puu	NOUN	S	...	5	nsubj	_	_
5 suudavad	suut	VERB	V	...	1	ccomp	_	_
6 talletada	talleta	VERB	V	...	5	ccomp	_	_
7 piisavalt	piisavalt	ADV	D	_	6	advmod	_	_
8 CO2	CO2	SYM	Y	...	6	obj	_	_

Figure 2: UD annotation of the sentence.

When converting the new corpus from CG to UD, we found that in addition to known problems in determining the function of clauses, it was also necessary to check determiners, names, copulas, elliptical constructions etc.

3.2 Determiners

Estonian CG annotation employs only pronoun part-of-speech, while UD also uses determiners. Although the transfer is mostly straightforward and lexicon based, there are some cases which only human could solve. In example (2), word-form *nende* can be determiner *these* or modifier *their*.

(2) *nende* *hindade* *puudumisel* ...
this-PL-GEN price-PL-GEN missing ...
they-PL-GEN price-PL-GEN missing ...

Missing of these/their prices ...

3.3 Names and appositions

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

[2] https://github.com/EstSyntax/EDT

annotates the last part of a multi-word name as the head. As for appositions, Estonian CG annotation scheme treats them as attributes. So, in example (3), the head of the name phrase is the rightmost node (comitative case of *Malouli*) in the CG annotation, and the leftmost (*finalist*) in UD.

(3) *Lahing Nordecon Openi finalisti*
battle Nordecon Open-GEN finalist-GEN
Laurent Malouliga
Laurent Malouli-COM

Battle with Laurent Malouli, the finalist of Nordecon Open

3.4 Copular constructions

The annotation of copula clauses is different in Estonian CG. Also, the definition of copula clause is wider as it is in CG and the straightforward rule-based conversion is not possible (Muischnek and Müürisep, 2017). CG annotation considers verb *be* as a root of the sentence (4), while it is a copula in the UD annotation. Figure 3 illustrates the differences of trees.

(4) *Homasho õige koht on*
Homasho-GEN right place be-3.SG
Goeido järel
Goeido-GEN after

Homasho's right place is after Goeido.

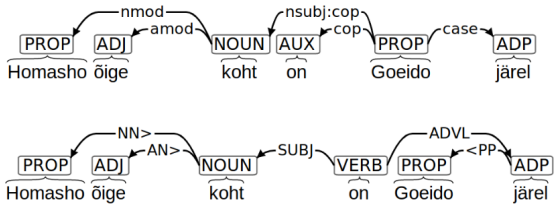

Figure 3: Copular constructions.

3.5 Elliptical constructions

CG-based treebank does not have any special label for ellipsis or orphan nodes. This annotation has been included into UD by special rule-based detector that can recognize some elliptical clauses but not all. As atypical elliptical clauses are quite frequent in the corpus of web language, they needed manual reannotation. Empty nodes have been included into UD syntax trees and the whole clause has an extra annotation of enhanced dependencies. In the sentence (5), the verb *pay* is omitted in the coordinated clause. Enhanced dependencies are marked as dotted arcs in figure 4.

(5) *ja siis jälle maksab mees ja siis*
and then again pay-3.SG man and then
jälle mina jne.
again I etc.

And then the man (husband) will pay and then I (will pay) again etc.

Figures 5 and 6 illustrate the format of CG and UD annotation of the EWTB sentence (5).

```
"<Ja>"
    "ja" L0 J crd CLB @J #1->4
"<siis>"
    "siis" L0 D @ADVL #2->4
"<jälle>"
    "jälle" L0 D @ADVL #3->4
"<maksab>"
    "maks" Lb V main indic pres ps3 sg ps af @FMV #4->0
"<mees>"
    "mees" L0 S com sg nom @SUBJ #5->4
"<ja>"
    "ja" L0 J crd CLB @J #6->9
"<siis>"
    "siis" L0 D @ADVL #7->9
"<jälle>"
    "jälle" L0 D @ADVL #8->9
"<mina>"
    "mina" L0 P pers ps1 sg nom @SUBJ #9->4
"<jne>"
    "jne" L0 Y adverbial @ADVL #10->9
"<.>"
    "." Z Fst #11->11
```

Figure 5: CG annotation of the sentence.

4 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).

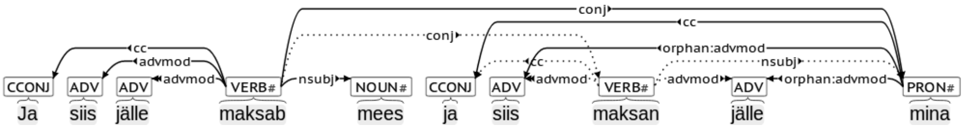

Figure 4: Elliptical constructions.

sent_id = ewtb2_149414_34
text = Ja siis jälle maksab mees ja siis jälle mina jne.

1	Ja	ja	CCONJ	J	_	4	cc	4:cc _	
2	siis	siis	ADV	D	_	4	advmod	4:advmod	_
3	jälle	jälle	ADV	D	_	4	advmod	4:advmod	_
4	maksab	maksma	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	0	root	0:root	_
5	mees	mees	NOUN	S	Case=Nom|Number=Sing	4	nsubj	4:nsubj	_
6	ja	ja	CCONJ	J	_	9	cc	7.1:cc	_
7	siis	siis	ADV	D	_	9	orphan:advmod	7.1:advmod	_
7.1	maksan	maksma	VERB	V	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	_	_	4:conj	_
8	jälle	jälle	ADV	D	_	9	orphan:advmod	7.1:advmod	_
9	mina	mina	PRON	P	Case=Nom|Number=Sing|Person=1|PronType=Prs	4	conj	7.1:nsubj	_
10	jne	jne	ADV	Y	Abbr=Yes	4	conj	4:conj	SpaceAfter=No
11	.	.	PUNCT	Z	_	4	punct	4:punct	_

Figure 6: UD annotation of the sentence.

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.