# Multiword Expressions in the Estonian Dependency Treebank

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen
University of Tartu
Estonia

*This paper is related to WG4 (Annotating MWEs in Treebanks).*

In the end of 2014, the first version of the Estonian Dependency Treebank (EDT) was created by manual review and reannotation of automatically analyzed text [1]. For this reason, the Treebank uses tags which are supported by current versions of Estonian syntactic analyzers. In particular, all syntactic annotation is done at the word level, using syntactic categories of standard grammar of written Estonian. Therefore, names for relations between subclauses are missing from annotation and multiword expressions are tagged partially. The only class of MWEs which is present in EDT are particle verbs (consisting of a verb and an adverb). Their automated detection rate is 97% [2].

The sentence '*Öö jooksul olid hundid kolm lammast maha murdnud*' (1) starts with an postpositional phrase *öö jooksul* 'during the night'. The verbal chain *olid maha murdnud* 'had killed' is split so the auxiliary olid 'had' occupies the second position in the clause and the rest of the verbal chain is situated at the end of the clause after the object *kolm lammast* 'three sheep', a typical word order of multiword predicates in Estonian. Main verb of the clause is a particle verb *maha murdma* 'kill down'; the particle *maha* 'down' functioning as a perfective marker. Figure 1 illustrates the annotation of example sentence (1).

(1) *Öö* *jooksul* *olid* *hundid* *kolm* *lammast* *maha* *murdnud*
night during be-AUX wolf-PL three sheep-PART down kill-PCP

'The wolves had killed three sheep during the night.'

Every sentence has been annotated at the morphological and syntactic level. The morphological description consists of the lemma, ending, POS, morphological information, and valency information. The syntactic description consists of a syntactic label and dependency information.

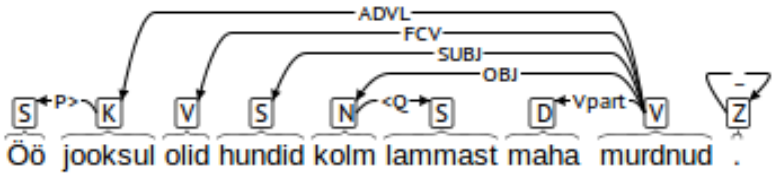


Figure 1: Dependency tree of sentence (1).

The fourth release of annotated treebanks in Universal Dependencies, v1.3 (`http://universaldependencies.org`), also contains Estonian treebank. The trees for Estonian were derived from the Estonian Dependency Treebank by automated conversion, using context-sensitive rules of Constraint Grammar [3]. Although both EDT and UD syntactic annotations are based on dependency grammar, they employ different sets of syntactic relations and analyse or annotate several linguistic phenomena (e.g. coordination, verbal chain) differently. During a manual review, the considerable amount of errors in treebanks have been fixed. These include errors revealed by syntactic validation queries (such as several uncoordinated subjects in one clause, incompatibility of morphological and syntactic annotation) and various technical errors. According to the UD annotation scheme, the particle verbs are tagged as 'compound:prt'.

Figure 2 illustrates UD tree of sentence (1).

The UD syntactic labels contain a separate set of labels for various multiword units and unanalyzable tokens (labels compound, mwe, goeswith, name and foreign). None of them is present in EDT annotation scheme.

Heuristic transformation rules allowed to recognize names and foreign phrases in EstUD treebank. The next challenge is to recognize other types of phrasal verbs as given in example (2).

(2) *Ma ei saa enam midagi aru, palun andke nõu*
I do not get anymore anything wit, please give advice

'I do not understand anything, please give an advice!'

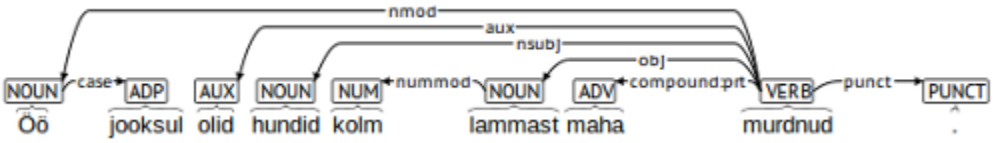


Figure 2: UD tree of sentence (1).

At the moment, these constructions have been analyzed as usual verbs with their arguments as regular objects.

# References


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

[2] Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. Estonian particle verbs and their syntactic analysis. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics: 6Th Language & Technology Conference Proceedings*, pages 338–342. Adam Mickiewicz University, 2013.

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