Multiword Expressions in the Estonian Dependency Treebank - WG4

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen

University of Tartu

Estonian Dependency Treebank (EDT)

- dependency-annotated treebank of written Estonian
- ca 400,000 words in ca 30,000 sentences
- genres: fiction, newspapers, scientific writing
- annotated semi-manually for
  - lemma
  - part of speech
  - morphological categories
  - syntactic functions
  - dependency relations
- freely available at `https://github.com/EstSyntax/EDT`
- reference: Muischnek et al., 2014

The original morphological tagset is a language-specific local standard, whereas the set of syntactic relations is based on Constraint Grammar (Karlsson, 1990; Karlsson et al., 1995) and coding of dependency relations is based on an expansion of Constraint Grammar (Bick & Didriksen, 2015).

(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.

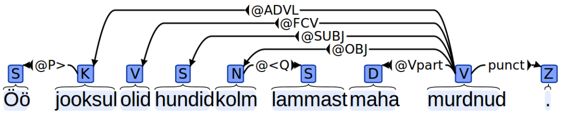

Figure 1: CG tree.

Particle verbs

Particle verbs are the only type of MWEs annotated in the original EDT.
A particle verb consists of a verb and a particle; the latter can often be homonymous with an adposition in Estonian, just like in English. In Estonian, just like in English a particle verb can be intransitive (Eng *take off*) or transitive (Eng *look it up*).
As for the particle placement and word order, the verb and the particle do not need to be adjacent to each other in Estonian and the order of the components may vary depending on the clause type, resembling much the behaviour of particle verbs in German.

(2) Ta **vaatas** üle aia ja nägi oma naabrit
s/he looked over fence-GEN and saw own neighbor-PART
S/he looked over the fence and saw his/her neighbor.

(3) Meelitustelt **läks** ta **üle** jõu kasutamisele
flattery-PL-ABL went s/he over force-GEN utilization-ALL
S/he switched from flattery to violence

Verb + Noun MWEs

Our plans for the future include annotating multiword expressions consisting of verb and noun.

- idioms

  (4) murrab pead
  breaks head-PART
  's/he thinks hard'

- collocations

  (5) hiivas ankru
  heaved anchor-GEN
  's/he heaved up anchor'

- support verb constructions

  (6) pidas kõne
  held speech-GEN
  's/he made a speech'

Automatic detection

Particle verbs have been annotated semiautomatically. Syntactic analyzer of Estonian detects ca 95% of particle verbs. 2-3% of particles were linked to wrong verb and 2-3% of particles have been annotated as part of particle verb erroneously (Muischnek et al., 2013).

- Rule based approach (Constraint Grammar).
- Altogether, the grammar for identification of Estonian particle verbs consists of approximately 500 rules and a thorough lexicon for 70 particles and corresponding lists of verbs.

Universal Dependencies

- `http://universaldependencies.org/`
- treebanks for 40 languages
- aimed at providing a cross-linguistically and typologically consistent inventory of categories and guidelines (e.g. Nivre, 2015)
- 234,000 words (in 18,100 sentences) from EDT were converted to UD and included in UD version 1.3

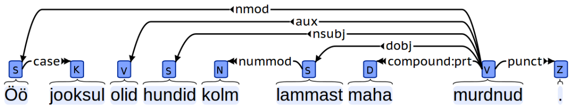

Figure 2: UD tree of sentence (1).

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 verbal MWEs as given in example.

(7) 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!

At the moment, these constructions have been analyzed as usual verbs with their arguments as regular objects.
Conversion process is described in detail in (Muischnek et al., 2016).

References

- Bick, E. and Didriksen, T. (2015). CG3 Beyond Classical Constraint Grammar. In *Proc. of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*, pp. 31–40.
- Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. *Proc. of Coling-90*. Vol. 3, pp. 168-173.
- Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.
- Muischnek, K., Müürisep, K., Puolakainen, T. (2016). Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Muischnek, K., Müürisep, K., Puolakainen, T., Aedmaa, E., Kirt, R., and Särg, D. (2014). Estonian Dependency Treebank and its annotation scheme. In V. Henrich et al., editors, *Proc. of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 285–291. University of Tübingen.
- Muischnek, K., Müürisep, K., Puolakainen (2013). Estonian Particle Verbs And Their Syntactic Analysis. In *Human Language Technologies as a Challenge for Computer Science and Linguistics: 6Th Language & Technology Conference Proceedings*. Poznan.
- Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pp. 3–16. Springer.

Acknowledgements

This work has been supported by Estonian Ministry of Education and Research (grant IUT 20–56 "Eesti keele arvutimudelid (Computational Models for Estonian)" and Norwegian-Estonian Research Cooperation Programme (grant EMP160 "SAMEST - Sami-Estonian language technology cooperation - similar languages, same technologies").