Automatic Detection of Estonian Argument Structure Constructions on the Example of Caused-Motion Verbs

Kertu Saul, Jelena Kallas, Kadri Muischnek

1 Introduction

This abstract presents the first results of a work in progress focused on automatically detecting argument structure constructions in Estonian on the basis of morpho-syntactically annotated corpora in the UD annotation schema. The research started with creating a preliminary workflow for the automatic detection of argument structure construction and testing it on caused-motion verbs. The results helped determine the primary problems associated with the task and come up with solutions to improve further work.

2 Background and motivation

Argument structure constructions are constructions that consist of a verb and its arguments (Goldberg 1995; Rätsep 1978: 15–18). While construction grammar also sees argument structure constructions as not being dependent on the meaning of the verb but rather the meaning of the clausal structure itself (Boas 2013: 235–236; Goldberg 1995: 224), this work is only concerned with detecting frequently used verb-specific argument structure construction and their fully-schematic forms. The example below exemplifies a variant of a fully schematic representation of a caused-motion construction consisting of a verb, a subject (causer), a direct object (entity being moved) and an oblique argument denoting destination.

nsubj V obj obl+ill
V = *riputama* 'to hang', *paiskama* 'to hurl', *toppima* ''to cram', *paigutama* 'to fit', *viima* 'to take/deliver', *tooma* 'to bring'

Ülikooli-d *too-vad*
university-PL.NOM bring-3PL
aju-d *Tartu-sse.*
brain-PL.NOM Tartu-ILL
'Universities bring brains into Tartu.'

Many languages already have corpus-based resources that describe argument structure constructions, e.g. FrameNet (Ruppenhofer et al., 2016; Ziem et al., 2019; Ohara et al., 2004; Dannélls et al., 2021; Torrent et al., 2014). Estonian has no such resource and creating one (semi-)automatically is the most realistic option. Previous similar work in Estonian was only done for automatically detecting verb-argument pairs, not whole constructions (Orasmaa 2013; Muischnek, Sahkai 2009).

Automatically detected argument structure constructions have both pedagogical and computational value. They can be used to supplement the grammatical information in the EKI Combined Dictionary (Langemets et al., 2021) by adding constructional information into its database. This work is also needed to improve and develop both the syntactic and semantic parsers for Estonian.

3 Methodology

The first step was compiling a test dataset for evaluating the quality of automatic detection. The initial material was collected from the manually morphologically and syntactically annotated UD-EDT corpus (UD version 2.12[1]), where each sentence occurring with one of the 28 caused-motion verbs in the sample was

[1] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5150

extracted for further manual analysis. Direct and abstract meanings were not differentiated. Frequently appearing argument structure constructions were manually identified for each verb and compiled in a dataset. The method resulted in 92 argument structure constructions where each member was represented by syntactic and morphological tags.

The next step was detecting argument structure constructions appearing with the same caused-motion verbs automatically. The data was collected from the Balanced Corpus of Estonian[2] consisting of 15 million words from 3 different genres. The corpus was automatically annotated both morphologically and syntactically in the UD framework.

The detection method is based on the co-occurrence frequency of a verb's direct dependents. The first step of this method was searching the corpus for each verb in the sample and counting each of its direct dependents' co-occurrence frequency with the verb. Two main restrictions were applied in this step: sentences with caused-motion verbs in valency-changing forms (e.g. impersonal voice) were not included and only those direct dependents that can be part of an argument structure construction (nsubj, obj, ccomp, xcomp, obl, advmod, compound:prt, case, non-finite advcl) were counted.

In order to better take into account the complex structure of the sentences and possible mistakes in automatic annotation, the frequency of every direct dependent's co-occurance with one another was counted in pairs aka bigrams. The bigrams were later combined into more complex verb-based argument structures. To be included in the construction, each dependent and pair of dependents had to appear in at least 5% of sentences with that verb. The bigrams were put together on the principle that if the bigrams a+b, b+c and a+c all crossed the 5% threshold, the sequence a+b+c also occurs. Finally, only the longest construction of each dependent was retained from these templates.

4 Results

A total of 107 verb-specific and 41 fully schematic argument structure constructions were automatically detected for 28 caused-motion verbs. The quality of detection was evaluated both in terms of individual dependents and complete constructions using precision and recall. 19% of arguments in the manually compiled material remained unidentified while 26% of automatically detected arguments were actually adjuncts. Out of the 19% of arguments left unidentified, 50% accounted for an oblique in the illative case, which is one of the many ways of expressing a GOAL in Estonian. Most of the dependents that were incorrectly identified as arguments were adverbial modifiers. Out of all argument structure constructions, 66% remained undetected while 33% of detected constructions were also in the test material. Out of the 66% of undetected constructions, 66% remained unidentified because of an incorrectly identified adverbial modifier, an unidentified oblique in illative or a missing subject.

Despite the low precision and recall, the method also identified previously undescribed argument structure constructions for two verbs: *liigutama* 'to move' and *pistma* 'to put/ to jab'. For *liigutama* the method found an argument structure construction with a subject, object and oblique in the inessive case representing LOCATION, which manifested itself only when the meaning of the verb was related to feelings.

Mingi ähmane aimdus liiguta-s
some vague hunch move-3SG.PST
end Luige-s.
itself.PART Luik-INE
'Some vague hunch moved itself in Luik.'

For *pistma* a construction with an object, an oblique in the additive case representing a GOAL and one in allative representing a RECIPIENT was identified.

Üks õde-de-st pist-i-s mu-lle
one sister-PL-ELA put-PST-3SG I-ALL

[2] https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en

käe püksi.
hand.GEN pants.ADDIT
'One of the sisters put a hand in my pants.'

5 Future work

This preliminary work was able to identify the main problems concerning automatic detection of argument structures.

1. It is difficult to automatically classify oblique and adverbial dependents into arguments and adjuncts;
2. Adjuncts that are frequent for all verbs should be differentiated from more verb-specific ones to be shown to L2 learners;
3. The argument structure constructions specific to a polysemic verb's less frequent meanings are not found;
4. Arguments can often be elliptic in corpus data and thus not identified;
5. Phraseological verbs are not identified as single lexical units in the source data;
6. There are too few example sentences in the gold standard UD-EDT corpus to make a comprehensive test dataset for less frequent verbs.

Future work will focus on tackling these various problems. In order to better differentiate arguments from adjuncts, adverbial dependents should be annotated with their semantic subclass. Polysemic verbs should be classified and annotated with their meaning using machine learning. To counter elliptic arguments, a default subject could be added to the constructions unless the verb is zero-valent. Phraseological verbs should be identified and set apart from the regular verb argument constructions. Test data should focus on verbs that have more data in the UD-EDT corpus and be annotated by several people.

6 Acknowledgements

This work was partly supported by the Estonian Research Council grant (PRG 1978).

References

Hans C. Boas. 2013. Cognitive Construction Grammar. *The Oxford Handbook of Constructional Grammar*, pages 233–252. Oxford University Press, Oxford.

Dana Dannélls, Lars Borin, Markus Forsberg, Karin Friberg Heppin, Maria Toporowska Gronostaj. 2021. Swedish Framenet. *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, pages 37–66. John Benjamins Publishing Company, Amsterdam.

Adele Goldberg. 1995. *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Margit Langemets, Kristina Koppel, Jelena Kallas, Arvi Tavast. 2021. Sõnastikukogust keeleportaaliks. *Keel Ja Kirjandus*, 64 (8–9): 755–770.

Kadri Muischnek, Heete Sahkai. 2009. Using collocation-finding methods to extract constructions and to estimate their productivity. In *Proceedings of the Workshop on extracting and using constructions in NLP*, pages 22−27.

Kyoko Ohara. 2018. Relations between frames and constructions: A proposal from the Japanese FrameNet constructicon. *Constructicography: Constructicon development across languages*, pages 141–165. John Benjamins Publishing Company, Amsterdam.

Siim Orasmaa. 2013. Verb Subcategorisation Acquisition for Estonian Based on Morphological Information. In *Proceedings of the 16th International Conference on Text, Speech and Dialogue*, pages 583–590.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. Institut für Deutsche Sprache, Mannheim.

Huno Rätsep. 1978. Eesti keele lihtlausete tüübid. *Eesti NSV Teaduste Akadeemia Emakeele Seltsi toimetised nr. 12*. Valgus, Tallinn.

Alexander Ziem, Johanna Flick, Phillip Sandkühler. 2019. The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, 35: 15–40.

Timponi Tiago Torrent, Ludmila Meireles Lage, Thais Fernandes Sampaio, Tatiane da Silva Tavares, Ely Edison da Silva Matos. 2014. Revisiting border conflicts between FrameNet and Construction Grammar: Annotation policies for the Brazilian Portuguese Constructicon. *Constructions and Frames*, 6 (1): 34–51.