Estonian copular and existential constructions as an UD annotation problem

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

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

Abstract

This article is about annotating clauses with nonverbal predication in version 2 of Estonian UD treebank. Three possible annotation schemas are discussed, among which separating existential clauses from copular clauses would be theoretically most sound but would need too much manual labor and could possibly yield inconcistent annotation. Therefore, a solution has been adapted which separates existential clauses consisting only of subject and (copular) verb *olema* be from all other *olema*-clauses.

1 Introduction

This paper discusses the annotation problems and research questions that came up during the annotation of Estonian copular sentences while developing Estonian Universal Dependencies treebank, especially while converting it from version 1 of UD annotation guidelines to version 2.

Copular clauses are a sentence type in which the contentful predicate is not a verb, but falls into some other category. In some languages there is no verbal element at all in these clauses; in other languages there is a verbal copula joining the subject and the non-verbal element (Mikkelsen, 2011, p. 1805).

In Estonian, there is mainly one verb, namely *olema* 'be', that functions as copula in copular clauses. Estonian descriptive grammar (Erelt et al., 1993) uses the term copular verb (Est 'köide') only for describing sentences with subject complements, stating that in such sentences the verb *olema* has only grammatical features of a predicate (time, mode, person). Also, the copula *olema* is semantically empty if used alone and it can not have any other dependents than non-verbal predicate. At the same time, verb *olema* is the most frequent verb of Estonian language which can also function as auxiliary verb in compound tenses, be part of phrasal verbs, and may occur in existential, possessor or cognizer sentences.

As the descriptive grammar of Estonian (Erelt et al., 1993) lacks more detailed treatment of copular sentences, the label "copula" has not been introduced into original Estonian Dependency Treebank (EDT) (Muischnek et al., 2014). In copular clauses *olema* is annotated as the root of the clause and other components of the sentence depend on it; that is also the case if the sentence contains a subject complement. As subject complements have a special label PRD (predicative) in EDT, such sentences can be easily searched.

Estonian treebank for UD v1.3 has been generated automatically from EDT using transfer rules. The guidelines for UD v1 implied that subject complements serve as roots in copular clauses. This analysis of copula constructions, according to UD v1 guidelines, extended to adpositional phrases and oblique nominals as long as they have a predicative function. By contrast, temporal and locative modifiers were treated as dependents on the existential verb 'be'.

Therefore, while converting EDT to UD v1, sentences with subject complements were relatively easily transferred to sentences with copular tree structure (Muischnek et al., 2016). However, oblique nominals and adpositional phrases were not annotated as instances of nonverbal predication.

Since UD v 2.0 assumes a more general annotation scheme for copular sentences, we faced several conversion problems and also linguistic questions. This paper provides insights into these research questions, gives an overview how copular clauses are annotated in some UD v2 treebanks for some other languages (Finnish, German, English) and describes what are the options for annotating Estonian sentences.

Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 79–85,
Gothenburg, Sweden, 22 May 2017.

In the remainder of this paper, Section 2 we give a short account of UD v2 guidelines for annotating copular clauses and show how these constructions are annotated in some UD v2 treebanks for Finnish, German and English. Section 3 is dedicated to copular constructions in Estonian language and Estonian UD versions 1 and 2. Some conclusions are drawn in Section 4.

2 UD annotation guidelines for nonverbal predication

According to the UD annotation scheme version 1, copular constructions are to be annotated differently from other clause types, analysing the predicative element as root and if there is an overt linking verb present, it should be attached to this nonverbal predicate as copula. The copula relation is restricted to function words whose sole function is to link a non-verbal predicate to its subject and which does not add any meaning other than grammaticalised TAME categories. Such an analysis is motivated by the fact that many languages often or always lack an overt copula, so annotation would be cross-linguistically consistent[1].

Version 2 of UD annotation guidelines extend the set of constructions that should be annotated as instances of nonverbal predication, defining six categories of nonverbal predication, namely those of equation, attribution, location, possession, benefaction and existence[2].

In order to get better overview of practical annotation of copular constructions cross-linguistically, we studied the v2 versions of UD treebanks of Finnish, which is the most closely related language to Estonian present in UD, and also German and English. As the language-independent annotation guidelines for UD version 2 were published in the very end of last year, there are no language-specific guidelines published yet. So we had to rely on treebank queries in order to gain information about annotating copular and related constructions in the afforementioned languages. We queried UD v2 treebanks using the SETS treebank search maintained by the University of Turku[3].

There are two UD treebanks for Finnish: the Finnish UD treebank, based on Turku Dependency Treebank, and Finnish-FTB (FinnTreeBank). In Finnish UD v2 treebank, clauses with *olla* 'be' are mostly regarded as instances of nonverbal predication, annotating *olla* as copula (1), among them also possessive clauses (2). However, if the clause contains only subject besides some form of *olla*, *olla* is annotated as root (3).

In Finnish FTB v2 treebank more clause types are annotated with *olla* as root, e.g. possessive clause (4) and clause containing predicative adverbial (5). It seems that annotation of copular constructions in Finnish FTB resembles that in version 1 of Estonian UD only subject complements are annotated as roots in copular constructions.

In the UD v2 treebank of German, sentences with subject complement are annotated as instances of nonverbal predication (6) and other instances of *sein* and *werden* 'be' seem to be annotated as main verbs, not copulas (7).

(1) Hyllyllä oli H&M
shelf-ADE was H&M
ROOT cop nmod
Home-tuotteita
Home-product-PL.PRT
nsubj:cop
'There were some H&M products on the shelf'

(2) Kuvia minulla ei ole
picture-PL.PRT I-ADE not be
nsubj:cop ROOT aux cop:own
'I have no pictures'

(3) Kun rahaa ei ole ...
If money-PRT not is
mark nsubj aux ROOT
'If there is no money'

(4) Meillä ei ole rahaa
we-ADE not is money-PRT
nmod:own aux ROOT nsubj
tuhlata.
waste-INF
acl
'We have no money to waste'

(5) Talonmies on juovuksissa.
Caretaker is drunkenness-INE
nsubj ROOT advmod
'The caretaker is drunk'

(6) Das Personal ist freundlich.
DET staff is friendly
det nsubj:cop cop ROOT

[1] http://universaldependencies.org/v2/copula.html

[2] http://universaldependencies.org/u/overview/simple-syntax.html#nonverbal-clauses

[3] http://bionlp-www.utu.fi/dep_search/

'The staff is friendly.'

(7) Ich war in dem Dezember bei
I was in DET December at
nsubj ROOT case obl case det
Küchen Walther.
Küchen Walther
obl flat
'I was in December at Küchen Walther.'

There are four English UD treebanks, but we queried only the largest of them, the English Web Treebank. It seems that predicative (e.g. Fig. 1) and locative (Fig. 2) constructions are analysed as instances of non-verbal predication, whereas existential clauses (Fig. 3) are not.

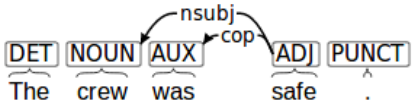

Figure 1: Predicative construction in the English UD treebank.

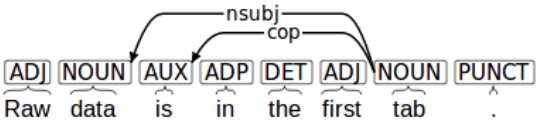

Figure 2: Locative construction in the English UD treebank.

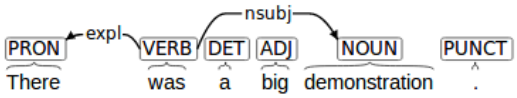

Figure 3: Existential clause in the English UD treebank.

As the above discussion illustrates, annotation of copular constructions varies across different languages and also across different treebanks. Having better documentation for v2 treebanks would facilitate better understanding of these differences. It would also be helpful for those teams which are still working on v2 of their treebanks to make more informed decisions.

3 Copular and existential constructions in Estonian

In Estonian, copular verb can not be omitted in normal writing or speech. However, there are some exceptions. First, copula is often omitted in headlines like (8).

(8) Valitsus otsustusvõimetu
Government indecisive
'Government is indecisive.'

And due to time pressure, copula, but also other verbs, can be omitted in online communication (9).

(9) Ma nii kurb
I so sad
'I am so sad.'

As a sidenote, although ellipsis of verb *olema* is rare in Estonian, it is still more frequent than in Finnish (Kehayov, 2008).

The annotation guidelines for version 2 of Universal Dependencies define six categories of nonverbal predication that can be found cross-linguistically (with or without a copula), namely equation (aka identification), attribution, location, possession, benefaction and existence.

As for Estonian, constructions expressing equation (10a), attribution (10b), location (10c), possession (10d), benefaction (10e) or existence (10f) are all coded using verb *olema*. In addition to the afforementioned clause types, there are also cognizer clauses (10g) that can be viewed as a subtype or metaphorical extension of the possessive clause type. Perhaps also quantification clause (10h) should be mentioned as a separate type.

(10) a. Mari on õpetaja.
Mari is teacher
'Mari is a teacher.'

b. Laps on väike.
Child is small
'Child is small.'

c. Laps on koolis.
Child is school-INE
'Child is at school.'

d. Lapsel on raamat.
Child-ADE is book
'Child has a book.'

e. See raamat on lapsele.
This book is child-ALL
'This book is for child.'

f. Oli tore kontsert.
was nice concert
'This was a nice concert.'

g. Lapsel on igav.
Child-ADE is boring.
'Child is bored.'

h. Lund on palju.
Snow-PRT is lot
'There is a lot of snow.'

Further in this section we will discuss three possible ways to annotate Estonian clauses containing the (copular) verb *olema*.

Among the constructions (10 a-h), existential clauses (f) are exceptional as they can consist also only of subject and some form of verb *olema*. In our opinion, such construction can not be annotated as an instance of nonverbal predication as there is no possible predicate except the verb *olema*. As for the sake of consistency, all existential clauses should be annotated in the same way, i.e. as instances of verbal predication, not nonverbal, we would have to distinguish existential constructions from all other *olema*-clauses.

In what follows in this Section, we will study if this solution can be applied while annotating real corpus sentences. For that we have to find out, if and how existential clauses can be identified. We start with investigating the linguistic features of the existential clause type on the background of main clause types in Estonian. Subsection 3.1 presents the linguistic features of Estonian existential constructions. Subsection 3.2 gives overview of constructions annotated as instances of nonverbal predication in version 1 of Estonian UD treebank. In subsection 3.3 we discuss three possibilities for annotating *olema*-clauses in Estonian UD v2 and conclude the section with adopting a compromise solution.

3.1 Existential clauses in Estonian

Characteristic features of Estonian existential clauses include the possibility of partitive subject (the default case of Estonian subject is nominative) and inverted word order (subject comes after verb), but existential clauses share these features with possessive clauses and also some other minor clause types. In order to understand the problem, we start with a small overview of main clause types in Estonian, explain how *olema*-clauses are distributed among those clause types, paying special attention to the existential clause type.

Descriptive grammars of Estonian (Erelt et al., 1993, pp. 14–15) and (Erelt, 2003, pp. 43–46) distinguish between three main clause types, depending on whether syntactic subject, semantic macrorole of actor and clause topic (theme) overlap, i.e. are coded by the same nominal. In so-called normal clauses, the same nominal functions as subject, actor and topic; in possessor-cognizer clauses, the possessed or cognized entity is both subject and topic, but not actor, which in turn denotes the possessor or cognizer. Subject noun denotes actor in existential clauses, but is rhematic.

Possessor-cognizer and existential clauses are regarded as marked clause types, also termed inverted clauses (Erelt, 2003, pp. 93–55), as they have inverted word order in pragmatically neutral sentences - XVS instead of SVX, otherwise typical for "ordinary" Estonian sentences. Subjects of these marked clause types can be in partitive case form, while nominative is the unmarked and statistically dominant subject case.

A few remarks about the possible case forms of subject in Estonian are in place here. Estonian descriptive grammars (Erelt et al., 1993, p. 15) and (Erelt, 2013, p. 36) state that existential and possessive clauses differ from other clause types in the possibility of subject case alternation: the subject is mostly in partitive case form in negative sentences (12), (14) and can be in partitive case form also in affirmative sentences (11),(13) if the referent of the subject noun is quantitatively unbounded.

(11) Selles klassis on targad lapsed / this-INE class-INE are smart-PL kid-PL / tarku lapsi.
smart-PL.PRT kid-PL.PRT
'There are smart kids in this class. / There are some smart kids in this class.'

(12) Selles klassis ei ole tarku lapsi.
this-INE class-INE not are smart-PL.PRT kid-PL.PRT
'There are no smart kids in this class.'

(13) Tal on head sõbrad / (S)he-ADE are good-PL friend-PL / häid sõpru.
good-PL.PRT friend-PL.PRT
'(S)he has good friends. / (S)he has some good friends.'

(14) Tal ei ole häid
(S)he-ADE not are good-PL.PRT

sõpru.
friend-PL.PRT
'(S)he has no good friends.'

Statistically, negation is the most powerful predictor of partitive subject (Miestamo, 2014). Of the clause types defined in UD documentation, existential clauses share the property of case-alternating subject with possessive clauses, but partitive subject is much less frequent in cognizer clauses, which can otherwise be seen as a constructional extension of the possessive clause. Partitive subject is the only option in quantification clause, regardless of its polarity, and in negative possessive clause.

Existential clauses differ from other marked clause types as they can consist also of subject only, besides the verb *olema*; in this case the clause merely states or negates the existence of an entity denoted by subject (15). There is also a special periphrastic verb form *olema olemas* (be be-INE) used for stating that something exists (16).

(15) On kontserte, kus loetakse
Are concert-PL.PRT where read-IMPS
ka luuletusi.
also poem-PL.PRT
'There are concerts where poems are also chanted.'

(16) Nõiad on olemas.
Witches are be-INE
'Witches exist.'

3.2 Nonverbal predication in version 1 of Estonian UD treebank

According to the first version of Universal Dependencies' guidelines, copular clauses consisting of a noun or an adjective, which takes a single argument with the subject relation were to be analysed as instances of nonverbal predication. The copula verb (if present) was attached to the predicate with the "cop" relation. This analysis of copula constructions was extended to adpositional phrases and oblique nominals as long as they had a predicative function. By contrast, temporal and locative modifiers were to be treated as dependents on the existential verb 'be'. So clauses containing some copular verb were to be divided between categories of verbal and nonverbal predication. In version 1 of Estonian UD treebank, only sentences containing verb *olema* and subject complement in nominative or partitive case form were annotated as instances of nonverbal predication (17). This was partly motivated by the fact that subject complements were already annotated using a special dependency relation (PRD) in our original treebank, the Estonian Dependency Treebank. All other copular clauses were annotated as instances of verbal predication, annotating form of verb *olema* as root (18).

(17) Mina olen Merlin.
I am Merlin
'I am Merlin'

(18) Ta on praegu kodus.
(s)he is now home-INE
'(S)he is at home now.'

3.3 Nonverbal predication in Estonian UD version 2: three possible solutions

As already mentioned in Section 2, version 2 of the Universal Dependencies' guidelines extends the number of constructions that fall into the category of nonverbal predication.

Among the Estonian copular constructions listed in the beginning of Section 3, existential constructions only stating or negating the existence of an entity expressed by subject and consisting only of verb *olema* and its subject (15) pose a problem for UD v2 annotation scheme. We are on the opinion that they can not be analysed as examples of nonverbal predication as one can not label subject as a predicate, so in these sentences verb form of *olema* 'be' has to be annotated as root, not as copula.

Thus we have three basic options for annotating copular constructions in Estonian UD v2: always annotate *olema* as copula; separate clauses consisting of verb *olema* and its subject from all other *olema*-clauses and, as third option, try to separate existential clauses from other *olema*-clauses. The fourth possible solution would be to stick to the solution we had in version 1 of Estonian UD, namely annotate only subject complements, i.e. nouns and adjectives in nominative or partitive case form as non-verbal predicates, but as this solution would violate the guidelines for UD version 2, we will not discuss it further.

We will take a closer look at the first three aforementioned options one by one.

Annotate all *olema*-clauses as instances of nonverbal predication

This would be the most straightforward solution from the point of view of UD v1 to v2 conversion process. The main drawback would be having to annotate subjects as predicates in clauses consisting only of verb *olema* and its subject (Fig. 4).

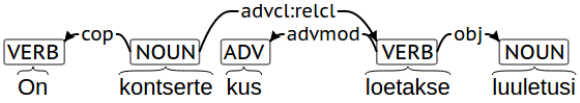

Figure 4: A subject as a root, annotation of the sentence (15).

Distinguishing subject-only clauses and other *olema*-clauses

Second possible option would be to distinguish two separate classes of *olema*-clauses basing on simple syntactic criterion: if the sentence consists only of some form of *olema* or periphrastic verb *olemas olema* and its subject, then *(olemas) olema* is annotated as root. All other sentences are annotated as instances of nonverbal predication. The distinction would be easy to make and the main drawback would be that existential sentences like (15) and (19) get different syntactic structures, which can be regarded as an inconsistent solution.

(19) Eile oli kontsert, kus
Yesterday was concert-PL where
loeti ka luuletusi.
read-IMPS also poem-PL.PRT
'There was a concert yesterday where poems were also chanted.'

Separate existentials and other *olema*-clauses

For the sake of consistency, all existential constructions should be annotated the same way, irrelevant whether they consist only of subject and verb or do they have more syntactic participants. This approach, although theoretically correct in our opinion, is not easy to apply in annotation practice.

As already described in subsection 3.1, Estonian existential clauses are defined as those, which subject is in partitive case in negative clauses and can be in partitive case also in affirmative clauses. In a pragmatically neutral affirmative clause, word order distinguishes between locative (20) and existential clauses (21). Negative variants of these clauses show the distinction - subject is in nominative case form in locative clause (22) and in partitive case form in existential clause (23). The example sentences are of course simplifications, especially with regard to word order; in the corpus sentences the word order does not distinguish existential clauses as it is determined mostly by information structure and depends heavily on larger textual context.

(20) Koer on aias.
Dog is garden-INE
'Dog is in the garden.'

(21) Aias on koer.
garden-INE is dog
'There is a dog in the garden.'

(22) Koer ei ole aias.
Dog not is garden-INE
'Dog is not in the garden.'

(23) Aias ei ole koera.
garden-INE not is dog-PRT
'There is no dog in the garden.'

So, if we would like to apply this solution, it would mean that human annotators have to go over all affirmative clauses that could possibly be existential clauses and make the distinction basing on their intuitions about the probable subject case in the negative counterpant of the affirmative clause under consideration - which means that there has to be more than one annotator for every clause. But Peep Nemvalts (2000), who has analysed Estonian existential sentences, comparing them with the same phenomenon in other languages, has concluded that it is impossible to distinguish Estonian existential sentences basing on formal criteria.

Therefore, after considering all possible solutions for distinguishing copular and non-copular usages of *olema* 'be', we had to make a compromise and adopt the second possible solution, i.e. distinguish subject-only clauses and other *olema*-clauses. In resulting annotated treebank, existential clauses are divided between instances of verbal and non-verbal predication, which can be regarded as a drawback. On the other hand, the resulting annotation is consistent, which is a clear advantage.

4 Conclusion

Universal Dependencies is planned to offer language-typologically relevant and cross-linguistically consistent annotation guidelines

for building dependency treebanks (Nivre et al., 2016). Its version 2, published only a few months ago, introduced a major change concerning annotating nonverbal predication: the repertoire of clauses that should be treated as examples of nonverbal predication was considerably broadened. Often real corpus data is a challenge even for well-premeditated theoretical constructs; even more so if this corpus data comes in more than 50 languages. So it should not be a surprise that there are still some open issues or inconsistencies.

This article tackled the problems concerning defining and annotating copular constructions in Estonian, with some brief cross-linguistic comparison.

We came forward with three possible annotation schemas, among which separating existential clauses from copular clauses would be theoretically most sound but would need too much manual labor and would possibly result in inconcistent annotation. So we will adapt the solution that, somewhat artificially, separates existential clauses consisting only of subject and (copular) verb *olema* from all other *olema*-clauses.

It seems that delimiting and annotating nonverbal predication and related phenomena is not entirely consistent cross-linguistically. In this article we had a look at a very small set of languages, but the analysis of nonverbal predication and copular constructions from a cross-linguistic (or cross-treebank) perspective deserves in-depth study. For the sake of better understanding the exact annotation of linguistic phenomena in different languages, thorough documentation of principles and decisions underlying the annotation would be beneficial.

As for Estonian UD treebank, the solution that at the first glance seemed most correct from the linguistic point of view, is (almost) impossible to achieve even by manual annotation.

Acknowledgements

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

Mati Erelt, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, and Silvi Vare. 1993. *Eesti keele grammatika II. Süntaks*. Eesti TA Keele ja Kirjanduse instituut.

Mati Erelt, editor. 2003. *Estonian Language. Linguistica Uralica Supplementary series*, volume 1. Estonian Academy Publishers, Tallinn.

Mati Erelt. 2013. *Eesti keele lausepetus. Sissejuhatus. Öeldis.* Tartu likool.

Petar Kehayov. 2008. Olema-verbi ellipsist eesti kirjakeeles. In *Emakeele Seltsi aastaraamat*, volume 54, pages 107–152. Eesti Teaduste Akadeemia Kirjastus.

Matti Miestamo. 2014. Partitives and negation: A cross-linguistic survey . In *Partitive cases and related categories*, volume 54 of *Empirical Approaches to Language Typology*, pages 63–86. De Gruyter Mouton.

Line Mikkelsen. 2011. Copular clauses. In Klaus von Heusinger Claudia Maienborn and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2, pages 1805–1829. Berlin: Mouton de Gruyter.

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014. 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.

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

Peep Nemvalts. 2000. *Aluse sisu ja vorm*. Eesti Keele Sihtasutus.

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).