WG4
Representation of MWEs in the Lithuanian Dependency Treebank

Jolanta Kovalevskaitė, Erika Rimkutė, Loïc Boizou
Vytautas Magnus University, Centre of Computational Linguistics

1. Introduction: Lithuanian Dependency Treebank

The Lithuanian Dependency Treebank (LDT) is a part of Clarin-LT infrastructure; the set period of working on LDT covers 2015-2016. The goal is to prepare 2300 sentences annotated according to the dependency grammar. The corpus itself consists of several text types: newspapers, journals, fiction (in each group approx. 690 sentences), and legal texts (approx. 230 sentences).

The guidelines for morphological annotation were taken from MULTEXT-East format (Erjavec 2012)[1]. Each part of speech is annotated using an individual set of morphological categories (from 2 to 14), e.g., verbal form *turi* (‘he/she has’), lemma *turėti* (‘to have’), annotation Vgmp3s--n--ni-; noun form *vertinimų* (‘evaluations’ in pl genitive), lemma *vertinimas* (‘evaluation’), annotation Ncmpgn-. Syntactic annotation follows a dependency model adapted from the Prague Dependency Treebank analytical layer (Hajič 1998), with some simplifications. The syntactic analysis is produced by a rule-based parser (Boizou et al. 2014). Both morphological and syntactic level are then corrected by linguists. Attempts to generate at least partial version of the Universal Dependencies[2] are discussed.

2. Types of MWEs in Lithuanian

In Lithuanian, there are several types of MWEs: nominal (named entities, idioms, collocations), verbal (idioms, collocations), proverbs. However, there is a particular type of MWEs of grammatical nature – these MWEs consist of two or more words (composed of inflective or uninflected parts of speech) and form semantically and syntactically unified, non-compositional unit that performs one syntactic function, e.g., multi-word adverbs, multi-word prepositions, multi-word particles, multi-word conjunctions, multi-word pronouns. MWEs from this group correspond to “MWEs of other categories” and “prepositional MWEs” in PARSEME annotation[3].

3. Representation of MWEs in LDT

As for the starting point, we annotate all words of different MWE types separately, except for those of grammatical nature.

3.1. MWEs of grammatical nature. These MWEs are identified automatically in LDT (from the list). The biggest group here are multi-word adverbs: *taip pat* (‘also, too’), *iš anksto* (‘in advance’), and multi-word pronouns: *kai kurie* (‘some’), *nė vienas* (‘any’). The groups of multi-word conjunctions (*vis dėlto* (‘however, nevertheless’)), multi-word particles (*vargu ar* (‘hardly’)), and multi-word prepositions (*iki pat* (‘to, until’)) are not numerous.

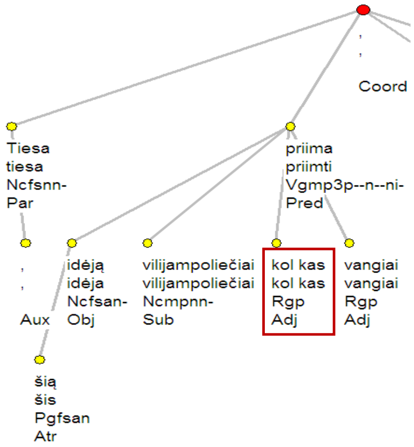

Fig. 1. Representation of the MWE *kol kas* ‘for the meantime’

[1] http://nl.ijs.si/ME/V4/msd/html/index.html.

[2] http://universaldependencies.github.io/docs/#language-

[3] http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

In LDT, all MWEs of grammatical nature are treated as single lexical units already on the morphological level, and appear as single nodes in the tree structures, e.g., MWE *kol kas* ('in the meantime, yet') is annotated as an adverb (Rgp) and functions as an adjunct (Adj) (see figure 1).

3.2. Collocations. Collocations are arbitrary, analyzable and flexible MWEs; their spectrum is wide, i.e. there are verbal, nominal, adverbial MWEs of different length (most of them are two- and three-word phrases), e.g., *priimti sprendimą* (‚to make a decision'), *atsakingas sprendimas* (‚responsible decision'). Analysis of their morphological and syntactical features reveals that some of them are fully flexible (both constituents can be declined – *atsakingas sprendimas* – adjective and noun in sg nominative, *atsakingo sprendimo* – adjective and noun in sg genitive), whereas others can contain only one fixed member (*priima sprendimą, priėmė sprendimą* – here only verb forms differ in tense and the noun remains in sg accusative) (for a full description of morphological and syntactical features of two-word collocations see Kovalevskaitė et al. 2015). There are also collocations which are used in one particular form, e.g., *pirminiais duomenimis* (literally: preliminary:INS.PL data:INS.PL, meaning 'preliminarily'). The flexibility of collocations allows to analyze each word of the collocation, and this way is quite reasonable bearing in mind that other words can be inserted into a collocation, because the Lithuanian word order is rather free.

3.3. Idioms. Idioms are analyzed by giving annotation to each of their constituents, but as these units are semantically non-compositional, it would be also useful to label them as an MWE. The MWE label could be additional, and provided after an idiom is analyzed syntactically. This is because there are idioms, especially with a verbal component, which are in some respect flexible: they appear in texts in particular morphological forms (Kovalevskaitė 2014; Kovalevskaitė et al. 2015), e.g., lemma *kasti karo kirvį* (‚to dig the hatchet') can appear as *iškas karo kirvį* (the underlined verb is 3st person, future). Usually, in case of verbal idioms, the form of a verb can differ in respect of tense, number, and person (for inflective forms), or can be used as a form of infinitive or participles (uninflected forms). These features of idioms provide an argument to analyze them by separate words. However, as in the case of collocations, there are also such idioms, which are fully frozen units (e.g., *vargais negalais* ‚with difficulty'), and in this respect more similar to the MWEs of grammatical nature.

3.4. Proverbs. Usually, proverbs are sentences which often appear as citations in texts, e.g., *kas ne su mumis – tas prieš mus* (‚he who is not with us is against us'). In the process of annotation, they are analyzed by separate words.

3.5. Multiword named entities. Named entities actually are also analyzed by separate words: names (*Valdas Adamkus*), geographical names (*Kauno rajonas*), names of companies, institutions (*Via Baltica*), etc.

4. Future work

The solutions for the representation of each type of MWEs are still under development, and more annotation scenarios have to be discussed, also from other languages.

The formal flexibility of the Lithuanian MWEs (mostly of collocations, idioms and named entities) and the rather free word order are important reasons to treat each word of these MWEs as a single syntactic node with its proper morphological and syntactic annotation. On the other hand, it is necessary to consider whether we have to apply more than one principle of annotation in respect to a particular MWE type; or, probably, it is more useful to annotate all words of different MWE types separately (except for those of grammatical nature) and to have a special label „MWE" for all MWEs. Then, we could document flexibility as well as frozenness of a particular MWE, and in LDT we would not need to deal with the lemmatization of MWEs, which, as research shows, is rather problematic (Boizou et al. 2015).

References

Boizou L., Kovalevskaitė J., Rimkutė E. Automatic Lemmatisation of Lithuanian MWEs. In *Proceedings of 20th Nordic Conference of Computational Linguistics NODALIDA 2015.* NEALT proceedings, vol. 23. Linköping, ACL anthology, 2015. http://aclweb.org/anthology/W/W15/W15-1808.pdf

Boizou L., Zamblera F. Syntactic Engine for the Lithuanian Language. In *Proceedings of the Sixth International Conference Baltic HLT 2014: Human Language Technologies – The Baltic Perspective.* Amsterdam, Berlin, Tokyo, Washington, DC, IOS Press, 2014, 69–75. http://ebooks.iospress.nl/publication/38006

Erjavec T. 2012: MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation* 46/1, 131-142.

Hajič J. 1998: Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: E. Hajičová (ed.): *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, Karolinum, Charles University Press, Prague, Czech Republic, 1998, 106-132. http://ufal.mff.cuni.cz/pdt2.0/publications/Hajic1998.pdf

Kovalevskaitė J. 2014: *Phraseme-type* and *Phraseme-token*: a Corpus-driven Evidence for Morphological Flexibility of Phrasemes. *Res Humanitariae*, XVI, 126–143. http://journals.ku.lt/index.php/RH/article/view/1016/1199

Kovalevskaitė J., Boizou L, Rimkutė E. 2015: Lietuvių kalbos dvižodžių junginių morfologinių ir sintaksinių ypatybių sąsajos. *Darbai ir dienos* 64 (in print).