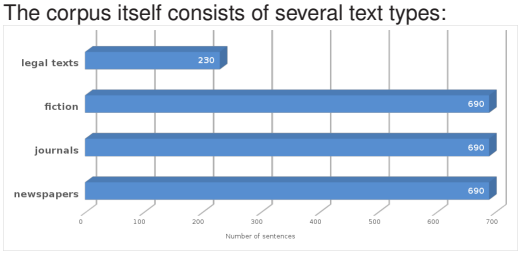
PARSEME 6th general meeting (7-8 April 2016, Struga, Macedonia)

The Representation of MWEs in the Lithuanian Dependency Treebank

Jolanta Kovalevskaitė, Erika Rimkutė, Loïc Boizou

Lithuanian Dependency Treebank (LDT) *Alksnis*

- 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:



LDT Annotations

Morphology

- According to MULTEXT-East format.
- Each part of speech is annotated using an individual set of morphological categories (from 2 to 14). Examples:
  - *turi* ('he/she has'), lemma *turėti* ('to have'), annotation Vgmp3s–n–ni- (verb, general, main form, present tense, 3rd person, singular, -gender, -voice, not negative, -definiteness, -case, not reflexive, indicative mood, -degree);
  - *vertinimų* ('evaluations'), lemma *vertinimas* ('evaluation'), annotation Ncmpgn- (noun, common, masculine, plural, genitive, -name).

Syntax

- Dependency model adapted from the Prague Dependency Treebank analytical layer.
- There are 18 syntactic functions (some of them combined, e.g. Pred_Atr, Pred_Co).

Pred (simple predicate)	Coord (coordination)
PredN (compound nominal predicate)	AuxC (subordinate conjunction)
PredV (compound verbal predicate)	AuxP (preposition)
Sub (subject)	AuxZ (emphasizing word)
Obj (object)	Aux, AuxK (other auxiliary)
Adj (adjunct)	
Atr (attribute)	

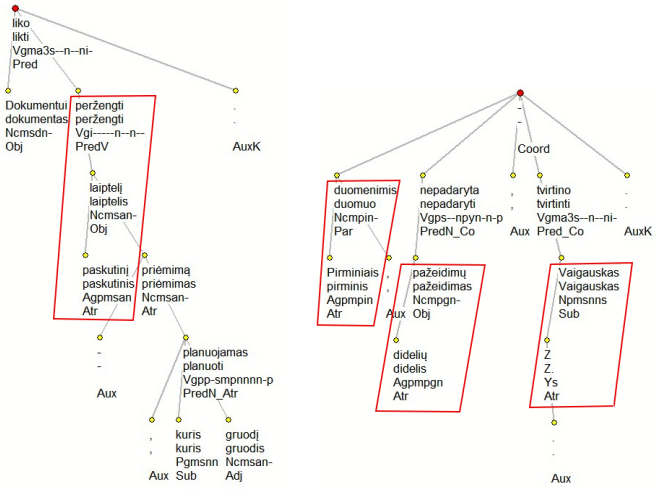
- The syntactic analysis is produced by a rule-based parser (Boizou and Zamblera 2014).
- Both morphological and syntactic annotations are performed automatically, but revised manually by linguists.

Types of MWEs in Lithuanian

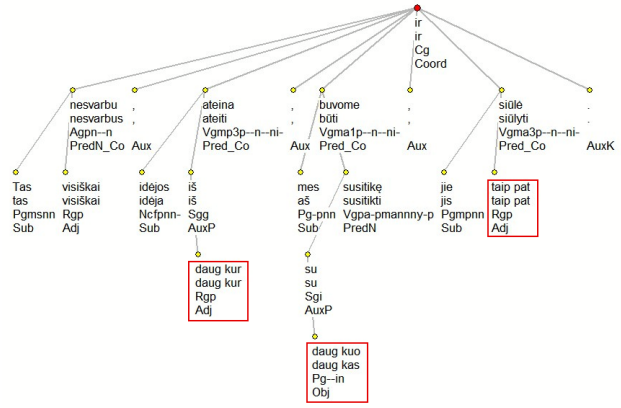
MWE type	Definition	Examples
Collocations	arbitrary, analyzable and flexible MWEs	**nominal collocation:** *atsakingas sprendimas* ('responsible decision') *pirminiais duomenimis* ('preliminarily' preliminary:INS.PL data:INS.PL) **verbal collocation:** *priimti sprendimą* ('to make a decision')
Idioms	semantically non-compositional MWEs	**nominal idiom:** *Achilo kulnas* ('Achilles heel') *vargais negalais* ('with difficulty') **verbal idiom:** *kasti karo kirvį* ('to dig the hatchet')
Proverbs	syntactically complete expressions with no slots to be filled	*kas ne su mumis — tas prieš mus* ('he who is not with us is against us')
Named entities	Geographical names, names of persons companies, institutions...	*Valdas Adamkus* *Kauno rajonas* *Via Baltica*
MWEs of grammatical nature	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	**multi-word adverbs:** *taip pat* ('also, too') *iš anksto* ('in advance') **multi-word pronouns:** *kai kurie* ('some'), *nė vienas* ('none') **multi-word particles:** *vargu ar* ('hardly') **multi-word prepositions:** *iki pat* ('to, until') **multi-word conjunctions:** *vis dėlto* ('however, nevertheless')

The Representation of MWEs Types in LDT

In LDT, we annotate all words of different MWE types separately, except for those of grammatical nature. 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.



MWEs of grammatical nature are treated as single lexical units already on the morphological level, and appear as single nodes in tree structures.



Future Works

- The solutions for the representation of each type of MWEs are still under development.
- Possible options:
  1. to apply more than one principle of annotation with respect to a particular MWE type;
  2. to have a special label *MWE* for all MWEs.

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
- Kovalevskaitė J., Boizou L, Rimkutė E. 2016: Lietuvių kalbos žodžių junginių morfologinių ir sintaksinių ypatybių sąsajos. *Darbai ir dienos* 64, 115–133.

Centre of Computational Linguistics
Vytautas Magnus University
http://tekstynas.vdu.lt
j.kovalevskaite@hmf.vdu.lt
e.rimkute@hmf.vdu.lt
l.boizou@hmf.vdu.lt