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:
  - newspaper
  - journal
  - fiction
  - legal

LDT Annotations

Morphology
- According to MULTTEXT-East format.
- Each part of speech is annotated using an individual set of morphological categories (from 2 to 14). Examples:
  - 'furške hasi', lemma 'furške' (‘to have’), annotation Vgmp3s–n–ni– ni (verb, general, main form, present tense, 3rd person, singular, -gender, -voice, not negative, -definiteness, -case, not reflexive, indicative mood, -degree).
- nominālum 
- 'ēs ļautam', 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).

Types of MWEs in Lithuanian

<table>
<thead>
<tr>
<th>MWE type</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collocations</td>
<td>arbitrary, analyzable and flexible MWEs</td>
<td>nominal collocation: aš tūkstos tarpynių (I have a hundred)</td>
</tr>
<tr>
<td>Idioms</td>
<td>semantically non-compositional MWEs</td>
<td>nominal idiom: ķipis pelēks (‘to be sad’)</td>
</tr>
<tr>
<td>Proverbs</td>
<td>syntactically complete expressions with no slots to be filled</td>
<td>verbal idiom: kā arī ir (‘and too’)</td>
</tr>
<tr>
<td>Named entities</td>
<td>geographical names, names of persons companies, institutions...</td>
<td>via Baltica</td>
</tr>
<tr>
<td>MWEs of grammatical nature</td>
<td>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</td>
<td>multi-word adverbs: pie būtinājumi (‘in advance’)</td>
</tr>
</tbody>
</table>

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