Syntactic analysis and syntactic patterns of MWEs

Agute Klints
Artificial Intelligence Laboratory (AiLab)
IMCS, University of Latvia

Keywords: Latvian language, multi-word expressions, MWE parsing, UD corpus

Tezaurs.lv - the largest Latvian electronic explanatory dictionary - contains 400,000 entries, among them 73,000 multi-word expressions (MWEs) (Grasmanis et al., 2023). Over the past two years, there has been an ongoing division of these MWEs into subgroups (proper names, multi-word terms, taxa, phraseological units, collocations). This classification provides additional information to the dictionary users regarding MWE functions within the language and promotes the study of phraseology in Latvian linguistics. Such grouping is beneficial for computational linguistics purposes as it facilitates such tasks as semantic parsing and information extraction, among others.

Most MWEs in the dictionary are linked to the corresponding word entries or a specific word sense that is included in the MWE. Therefore, dictionary users can either search a specific expression or find it in the matching word entry.

While this task focuses on the semantic sorting of MWEs that are included in the Tēzaurs.lv data, it is planned that we will also conduct a pilot study which will focus on syntactic analysis and patterns, considering the lexical and frame semantic aspects and patterns of a selected subset of MWEs.

The syntactic analysis and annotation of MWEs will be supported by an automatic dependency parser. The automatic parse trees will be manually post-edited, and the internal structure of MWEs will be examined and enriched with additional annotations. Such annotations will help Tēzaurs.lv users to find MWEs given in different grammatical cases (currently, MWEs can be found only by providing the base form) and will provide the users with inflectional tables of MWEs. It will also provide the ability for parser to automatically identify MWE in text.

We have intended that for a small group of MWEs parsing would be relatively easy as, for example, multi-word proper names and terms can be found on databases and are invariable. However MWEs that are not fixed and are variable (phraseological units, collocations) are not that predictable, as word order, lexical components, syntactic structure, word cases etc. can vary in many ways. This is the problem that syntactic parsing would help to solve.

At the moment we have recorded MWEs of Latvian lexicon in an electronic explanatory dictionary, we have manually semantically sorted them. Next step is planned to be the creation of Latvian MWE corpus with additional information about each MWE, their nature, inflections, varying, manual MWE marking in text, and MWE structure parsing.

MWE processing in Latvian language is in its beginning phase, and it is planned to analyze MWEs on the UD level, defining morpfology, principal forms, MWE syntactic structure, syntactic dependency etc. Therefore the UniDive training school course will be useful for reviewing and comparing our parsing guidelines with existing ones. Besides, the Latvian language already has a UD corpus (without annotated MWE layer), and adding a new annotation layer (MWEs) on top of our UD corpus could be another step in improving the existing corpora.

References:

1. M. Grasmanis, P. Paikens, L. Pretkalnina, L. Rituma, L. Strankale, A. Znotins, and N. Gruzitis. 2023. Tezaurs.lv – the experience of building a multifunctional lexical resource. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*, p. 400–418. Lexical Computing CZ s.r.o.