The goal is to build a huge computational lexicon for Macedonian language, which will enable recognition and tagging of all inflectional forms of MWEs.

The workflow:
1. Extract potential MWEs from a huge corpus
2. Filter them using a NLP tool
3. Manually polish them
4. Classify them
5. Develop inflectional classes and assign them to obtained MWEs

Towards a New Computational MWE Lexicon
What has been achieved so far?

1. 400,000+ potential MWEs have been extracted from Wikipedia’s titles

2. 80,000+ passed the first filter (foreign alphabets, numbers)

Towards a New Computational MWE Lexicon
What is to be done?

2. Filter them additionally using various syntactic structures (AdjN, NpN, NpAdjN, AdjAdjN, AdjNpN, AdjNAdjN, NcN, NN, AAdvN) and an existing morphological lexicon

3-5. Manually polish them, classify them, develop inflectional classes and assign them to obtained MWEs

Example of a lexical entry:

ад хок, ADV+FLX=IMENO+Lxc+Fxd

ад хок - MWE (eng. ad hoc)
ADV – Grammatical category
IMENO – Inflectional class
Lxc – Lexical idiomaticity
Fxd – Fixed expression

Towards a New Computational MWE Lexicon