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

What has been achieved so far?

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

394939	Револуција на црвените каранфили
394940	Рибино масло
394941	Категорија:Родени во Хајделберг
394942	Ханс-Јоахим Хопе
394943	Еуростандард банка
394944	Оддел за економски и социјални работи на Обединетите нации
394945	Службени јазици на Обединетите Нации
394946	Економија на Германија
394947	Буџетот на Европската Унија
394948	Фондацијата на Обединетите Нации
394949	Германски претседател
394950	Седиштето на Обединетите нации
394951	Вештачка интелигенција
394952	Manuel Neuer
394953	Индијанците во САД

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

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