Macedonian language, like other South Slavic languages, is rich with multiword expressions (MWEs). There are several printed dictionaries related to MWEs, comprising between 5,000 and 20,000 phrases.

A huge computational lexicon, dealing with MWEs, does not exist. There is only a small morphological lexicon of compound words (700+ entries).

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

A huge morphological lexicon of simple words (85,000 lemmas, 380,000 word forms) is on disposal, which will be used for creating syntactic grammars.

What has been achieved so far?
1. Wikipedia was used as a source of potential MWEs. An application was developed to extract all titles and subtitles. As a result, more than 400,000 potential MWEs were obtained.

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

   The obtained list of potential MWEs still contains a lot of non-useful items, which are to be filtered additionally.

What is to be done?
2. Filter the obtained MWEs 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.

   It is expected that several thousands MWEs will remain. In order to gather more, other techniques should be used.

Example of a lexical entry:

ад хок, ADV+FLX=IMENO+Lx+c+Fxd

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