

Towards a New Computational MWE Lexicon

Aleksandar Petrovski, Katerina Zdravkova
Working group 1: Lexicon-Grammar Interface

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 existing morphological lexicon of compound words

Dictionary contains 704 entries

```
а виста, ADV+FLX=IMENO+UNAMB
а konto, ADV+FLX=IMENO+UNAMB
а ла карт, ADV+FLX=IMENO+UNAMB
а ла, ADV+FLX=IMENO+UNAMB
а проп, ADV+FLX=IMENO+UNAMB
Авирени штати, N+FLX=IMENO+UNAMB
Австро-Унгарија, N+FLX=IMENO+UNAMB
Австро-Унгарија, N+FLX=IMENO+UNAMB
авторски табан, N+FLX=IMENO+UNAMB
аграрен професионализам, N+FLX=IMENO+UNAMB
аграрна реформа, N+FLX=IMENO+UNAMB
аграрна реформа, N+FLX=IMENO+UNAMB
ад акта, ADV+FLX=IMENO+UNAMB
ад хок, ADV+FLX=IMENO+UNAMB
адверзивна реченица, N+FLX=IMENO+UNAMB
академска расправа, N+FLX=IMENO+UNAMB
академски трафанин, N+FLX=IMENO+UNAMB
акционен радиус, N+FLX=IMENO+UNAMB
Ал фатак, N+FLX=IMENO+UNAMB
алај-бајрам, N+FLX=IMENO+UNAMB
алај-бет, N+FLX=IMENO+UNAMB
алеа јахта ест, INT+FLX=IMENO+UNAMB
ал-пари, ADV+FLX=IMENO+UNAMB
алтер ето, PRO+FLX=IMENO+UNAMB
амбер-бој, N+FLX=IMENO+UNAMB
ан блок, ADV+FLX=IMENO+UNAMB
ан генерал, ADV+FLX=IMENO+UNAMB
ан мас, ADV+FLX=IMENO+UNAMB
ан пасан, ADV+FLX=IMENO+UNAMB
```

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.

The workflow:

Extract potential MWEs from a huge corpus

Filter them using a NLP tool

Manually polish them

Classify them

Develop inflectional classes and assign them to obtained MWEs

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.

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

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+Lxc+Fxd

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