Towards a New Computational MWE Lexicon

Aleksandar Petrovski, Katerina Zdravkova

Working group 1

Macedonian language, like other South Slavic languages, is rich with multiword expressions (MWEs). There are several printed dictionaries related to MWEs. Velkovska is an author of a notebook dealing with Macedonian phraseology [1] and a dictionary, which consists of 5,000 phrases [2]. Another printed dictionary, by Širilov and Dimitrovski, is published with 20,000 phrases [3].

Computational resources, dealing with MWEs, don't exist. The goal of this work is to build a huge computational lexicon which will enable the recognision and tagging of all inflectional forms of MWEs. The workflow that will be followed to achieve this task is:

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

Wikipedia will serve as a corpus for MWE extraction. With 300,000+ articles, it is definitely a huge source. The extraction of MWE will be done using a parser which determines the potential combinations of words organized into blocks of words with variable length. The frequency of their appearance will be assigned to isolate the most frequent blocks, which are the most valuable for further research.

The NLP tool which will be used for filtering is NooJ. It is a linguistic development environment that allows users to build large-coverage, mostly finite state descriptions of natural language and apply them to large texts [4]. It integrates morphology and syntax, thus enabling morphological operations inside syntactic grammars. Number of syntactic grammars will be developed for filtering. A huge computational lexicon already exists for NooJ [5], containing 85000+ single word entries. This resource is a necessary prerequisite for creating syntactic grammars. They are supposed to reflect potential syntactic structures for MWEs, which could be classified into 4 groups:

1. Nominal
2. Verbal
3. Adjectival
4. Prepositional

Vast majority of MWEs will belong to the nominal group. Various syntactic structures will be used, for example: AdjN (*tvrda glava*), NpN (*raka na srce*), NpAdjN (*fond za zdravstveno osiguruvanje*), AdjAdjN (*evropski pretpristapni fondovi*), AdjNpN (*mrtva bukva na hartija*), AdjNAdjN (*bistra voda, mirna glava*), NcN (*luk i voda*), NN (*bure barut*), AAdvN (*pijan kako zemja*). Many other structures are possible for all other groups.

It is expected that after the filtering, a huge number of non-MWEs will be in the list [6] and the manual polishing will be necessary. The criteria defined in [7], which include lexical, syntactic, semantic, pragmatic and statistical idiomaticity, will be applied.

The classification of MWEs into lexicalised phrases and institutionalised phrases will be used when classifying obtained MWEs [7]. Lexicalised phrases are MWEs with lexical, syntactic, semantic or pragmatic idiomaticity. They can be further split into fixed expressions, semi-fixed expressions and syntactically-flexible expressions. Syntactically-flexible expressions are not candidates for lexical entries, since they have not fixed word order and decomposition and should be presented by syntactical grammars.

A lexical entry will comprise a MWE, its grammatical category, its inflectional class, its group and its MWE class. Inflectional classes which describe the inflectional behavior of MWEs, expressed by regular expressions, will be built. The intention is to build a lexicon which will be easily transformable to different formats. For that purpose, additional computer applications will be developed.

References:

1. Velkovska, S. "Beleški za makedonskata frazeologija", Institut za makedonski jazik Krste Misirkov, 2002, ISBN 9989-640-39-4
2. Velkovska, S. "Makedonska frazeologija so mal frazeološki rečnik", 2008, ISBN 978-9989-57-599-0
3. Širilov, T., Dimitrovski, T. "Frazeološki rečnik na makedonskiot jazik", Vol.1-3, Ogledalo 2009, ISBN 9989686130, 9789989686139
4. Silberstein, M. „NooJ Manual", Available at http://www.nooj4nlp.net/
5. Petrovski, A. "Morfološki kompjuterski rečnik – pridones kon makedonskite jazični resursi", PhD thesis, University St. Cyril and Methodius, Faculty of natural sciences and mathematics, Institute of informatics, Skopje 2008
6. Bekavac, B.; Tadić, M. "A Generic Method for Multi Word Extraction from Wikipedia" Proceeings of the ITI 2008 30th Int. Conf. on Information Technology Interfaces, 2008.
7. Baldwin, Timothy and Kim, Su Nam Kim "Multiword Expressions" In N. Indurkhya and F. J. Damerau (Eds.), Handbook of Natural Language Processing (2 ed.), pp. 267-292, 2010