How to create a MWE lexical entry?

WG1

Aleksandar Petrovski
International Slavic University
Faculty of Informatics
Sveti Nikole, Macedonia
a.petrovski.sise@gmail.com

Katerina Zdravkova
University Sts Cyril and Methodius
Faculty of Computer Science and Engineering
Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

Abstract

One suggestion how to create a MWE lexical entry is given. Although NooJ syntax has been used, it is applicable to any other NLP system.

Classification adopted by Parseme action has been used when defining codes of MWE properties. Problems with lexicalising flexible MWEs are discussed.

1 Introduction

When a MWE lexicon is developed, it is crucially important to develop a classification which captures the general properties of MWE classes, but at the same time allows for the encoding of information particular to a given MWE instance /1/. Here we use the classification presented by Baldwin and Kim /1/. MWEs are classified by three main properties: flexibility, syntactic structure, and idiomaticity.

Since NooJ will be used as a NLP tool, NooJ syntax for lexical entries is presented first /2/. Then, codes reflecting the classification from /1/ are suggested. These codes will be used when a MWE lexicon for Macedonian is built. In the end, challenges using a NooJ MWE lexicon in parsing process are considered.

Although examples presented are from the Macedonian lexicon, considerations made are applicable for other languages.

2 NooJ syntax

A NooJ lexicon consists of lexical words, usually presented in a form similar to that used in common lexicographic practice.

These words are accompanied by a code that describes the inflectional behavior of the simple word. The structure of a NooJ lexical entry is presented by the following scheme:

$$W_{lex},K_{pos}+FLX=K_n+K_{kat} \quad (1)$$

- W_{lex} is a simple lexical form which is usually represented in a form of lexical entry;
- K_{pos} is a code representing the grammatical category of the word (PoS – part of speech);
- K_n is the class code that describes the grammatical properties of W_{lex}, i.e. its inflectional paradigm. This code goes always right to the key word FLX;
- K_{kat} are additional grammatical codes.

In case of a MWE lexicon, W_{lex} will be a MWE in its canonical form, K_{pos} its grammatical category, K_n will be the class code which describes the inflectional paradigm of the MWE and K_{kat} will be the codes presented in this paper.

3 MWE lexical entry

All MWEs are classified by three properties: fixedness/flexibility, idiomaticity, and syntactic structure. The classification, along with assigned codes, is given in Table 1.

Adverbial syntactic structure has been added to the original classification since it exists in the Macedonian language (e.g. *горе-долу* eng. up and down).

Deeper syntactical classification, like different NPs, light and phrasal verb structures etc, are beyond the scope of this paper.

Property	Value	Code
Fixedness/Flexibility	Fixed	FxdF
	Semi-fixed	SFxF
	Flexible	FlxF
Idiomaticity	Lexical	LexI
	Syntactic	SynI
	Semantic	SemI
	Pragmatic	PrgI
	Statistical	StaI
Syntactic structure	Nominal	NomS
	Verbal	VerS
	Adjectival	AdjS
	Adverbial	AdvS

Table **1**: Classification of MWEs with codes

Using the codes from Table 1, a lexical entry in a NooJ MWE lexicon will look like:

кабелска телевизија,N+FLX=N2
+SFxF+PrgI+StaI+NomS (2)

where *кабелска телевизија* is a MWE (eng. cable television), *N* stands for noun - grammatical category of the MWE, *N2* is the inflectional class code that describes the inflectional paradigm, *SFxF*, *PrgI*, *StaI* and *NomS* are codes from Table 1, describing MWE's properties.

Since N2 describes the lexical entry's inflectional behaviour, NooJ is able to recognise all of its 9 inflectional forms /3/.

4 Challenges

The MWE from (2) is an example of a semi-fixed MWE. NooJ formalism allows lexicalisation of these types of MWEs, since all inflectional forms are recognised. The situation with fixed MWEs is even simpler. But, what should be done with flexible MWEs? Let's have a look at a possible lexicon entry of a verbal MWE:

шири гласини,V+FLX=V2
+FlxF+SemI+VerS (3)

This is an example of a verbal flexible MWE (eng. spread rumours). Inflectional class V2 will comprise all inflectional forms of the lexical entry. But, this MWE could appear in different word orders, e.g. *гласини се шират* – reflexive form or *гласини се ширени* – passive form. Internal modifications (*мој шири опасни гласини* eng. he spreads dangerous rumours) and extractions (*какви гласини мој шири?* eng. what kind of rumours does he spread?) are also possible. Combinations of e.g. internal modification and passivisation make the situation even worse.

MWEs with non verbal syntactic structures could also be flexible:

црн како јаглен,A+FLX=A2
+FlxF+SemI+AdjS (4)

This is an example of an adjectival MWE (eng. as black as coal). It is coded as flexible, since internal modifications are possible (*црн како најцрн јаглен* eng. as black as blackest coal).

Some of these forms might be captured by inflectional paradigms (e.g. passive forms), but generally speaking, using a NooJ MWE lexicon as a single resource is useless in these cases. The only way to recognise all forms of flexible MWEs is building syntactical grammars. One grammar should be built for each flexible MWE, which will comprise passivisation, internal modification, and extraction.

5 Conclusion

A possible MWE lexical entry of a NooJ lexicon is presented. High level coding of MWE's properties is proposed. NooJ lexicons are very good when dealing with fixed and semi-flexed MWEs. For flexible MWEs, due to possible change of word order, lexicons can't help much. The only way is to build a syntactical grammar for each MWE.

It would be interesting to analyse how other NLP tools deal with flexible MWEs and to compare them with NooJ.

References

1. 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
2. Silberstein, M. „NooJ Manual", Available at http://www.nooj4nlp.net/
3. 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