Dictionary of Slovak Multiword Expressions

Ján Genči[1], Martin Ološtiak[2] and Ján Staš[1]
[1]Technical University of Košice, Slovakia
[2]University of Prešov, Slovakia

Relevant to WG1

The poster presents results of the three-year project entitled "Slovak multiword expressions - lexicographical, lexicological and comparative research", funded by the Slovak Research and Development Agency in 2012-2015 years.

Due to its linguistic nature, the construction of the dictionary was planned to be done by manual extraction of MWEs from various sources:

- entries of basic Slovak monolingual dictionaries;
- diploma theses;
- manual excerption from various sources - thesauri, translational and terminological dictionaries, encyclopedias, elementary school textbooks.

Initially, about 75,000 MWEs were collected by the linguists from the mentioned sources. After duplicate elimination, there remained more than 56,500 unique MWE entries. The frequency of occurrence in the database of the Slovak National Corpus was determined for every unique entry and then about 14,000 the most frequent entries were selected for the publication in the two-volume Dictionary of the Slovak Multiword Expressions. The first volume contains MWE entries organized by so-called hyper-entries (content word or „main" word from the MWE) and detailed description with translation of particular MWEs to the English, French, German, Russian and Spanish languages (see Table 1). The second volume contains the Slovak and several foreign language indices, and reference entries to facilitate search of corresponding entries in the first volume.

Computer based processing of MWEs

MWE entries have been frequency-wise statistically processed against the Slovak National Corpus (summarized in the Table 3) including all word forms of particular MWE. For this purpose, a custom database of full word forms were developed and applied for computation of frequencies based on regular expression data selection. Some approaches regarding the quality of the data (excerpted and full word forms database) and reduction of computational complexity was also considered.

Table 1. Examples of dictionary entries with hyper-entries (volume 1)

INTERVIEW

exkluzívne interview [intervjú] s. -neho neskl. publ. rozhovor určený výlučne istému médiu; syn. exkluzívny rozhovor: *poskytnúť denníku exkluzívne interview*

A exclusive interview; **F** entretien exclusif; **N** das Exklusiv-Interview; **R** эксклюзивное интервью; **Š** entrevista exclusiva

LEKÁR

zubný lekár m. -ného -ra kto odborne ošetruje zuby; dentista, stomatológ: *ambulancia zubného lekára* ■ **zubná lekárka** ž. -nej -ky ♦ hovor. **zubár** m. -ra

A dentist; **F** dentiste; **N** der Zahnarzt; **R** врач-стоматолог / зубной врач; **Š** odontólogo

Table 2. Examples of reference entries (volume 2)

čiastočný ↗ dôchodok, invalid, úhrada, úväzok, zatmenie
minerálny ↗ látka, olej, soľ, prameň, voda
španielsky ↗ čižma, gitara, chrípka, jazyk, stena, vtáčik

act of mercy	skutok milosrdenstva
action comedy	akčná komédia
action hero	akčný hrdina
action movie	akčný film
action novel	akčný román

Table 3. Frequency of occurence of the n-gram types in the excerpted data

***n*-gram structure of entries**	**frequency of occurence**
2-gram	81 993
3-gram	6 491
4-gram	948
5-gram	173
6-gram	29
7-gram	6
8-gram	4
9-gram	1

Conclusion

The presented Dictionary of the Slovak Multiword Expressions was processed manually, except determining of some statistical data related to statistics obtained from the database of the Slovak National Corpus (SNC). In the future work, we plan to process the SNC data to determine the MWEs directly and compare results of manual and statistical processing.