Dictionary of Slovak Collocations*

Peter Ďurčo[1], Radovan Garabík[2], Daniela Majchráková[2], and Matej Ďurčo[3]

[1] Univerzita sv. Cyrila a Metoda v Trnave, Trnava
[2] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[3] Austrian Academy Corpus, Austrian Academy of Sciences, Vienna

Abstract. Presented lexical database of Slovak language collocation should cover collocation profiles of several hundred words of different parts of speech (nouns in the first phase of the project) and will be a base of a modern collocation dictionary. The database is built using MediaWiki engine, which offers excellent remote collaboration features along with automatized processing possibilities.

1 Introduction

The standard use of corpora for linguistic research and lexicography is aimed predominantly at the examination of occurrences and co-occurrences of word forms and lemmata. The main goal is to acquire data about semantic, grammatical and combinatorial behavior of words.

For the Slovak language, the only one existing collocation dictionary has been published in 1931, with a revised edition in 1933 (the author called this book 'a dictionary of phrasemes', but in fact it has been a dictionary of not only phrasemes, but also of common word collocations) [13, 14]. Clearly, since then the whole language underwent immense changes in almost all of its parts, starting with the whole sociolinguistic situation and ending with substantial changes in the vocabulary and orthography. By today, the dictionary is mostly of diachronic importance, and there is a notable gap in Slovak language lexicography concerning a database of collocations – modern approaches in lexicography, especially the use of large language corpora fill the gap somewhat, but they still cannot replace a well documented, systematically built dictionary.

Presented electronic dictionary of Slovak collocations is being compiled at the University of St. Cyril and Methodius, Trnava in cooperation with the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava. The project on Slovak collocations that started in 2007 is the first of its kind in Slovakia and is aimed at the registration and description of selected multiword lexemes and phrasemes as well as typical collocations with restricted collocability. The dictionary provides an overview of the combinatorial behaviour of words, in the first phase the most frequent nouns extracted from the Slovak National Corpus database, with the intention to include also verbs, adjectives, adverbs and particles. Currently, the database contains information about nouns and (as a separate subproject) particles. The combinatorial potentials of word forms of a word are the basis for the creation of so-called collocational templates which the patterns of collocations are based on [17]. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partical collocations.

2 Obtaining collocation profiles

An efficient tool for modelling semantic proximity of words and their collocation profiles in large lemmatized corpora is the sketch engine[4] [10] – a corpus tool which generates word sketches, i. e. corpus based summaries of a word's grammatical and collocational behaviour. Disadvantages of the

* The lexical database has been supported by the grant agreement VEGA 1/0006/08 *Konfrontačný výskum kolokácií v slovenčine a v nemčine*. The study and preparation of these results have been partly supported by the EC's Seventh Framework Programme [FP7/2007-2013] under the grant agreement 211938 MONDILEX.

[4] `http://www.sketchengine.co.uk/`

sketch engine are long lists of isolated lemmata and too many automatically generated redundant data in the results, obtained through fixed set of unary, dual, symmetric and trinary rules, which do not always correspond to natural collocational clusters in the language.

The basic tool for searching collocations for each entry is the corpus manager client Bonito which provides searching, sorting and statistical evaluation of collocations. By using this tool we can observe each given word, extract concordances for each word to get an overview of its behaviour in a context, get statistical information like absolute frequency, MI-score, t-score, MI-score, MI3, log likelihood, min. sensitivity and salience to recognize word co-occurrences [11].

Despite these new language technological analysis, scepticism still prevails regarding the possibility of seizing and of describing examined data completely. This scepticism results particularly from two problems. Word co-occurrences represent a diffuse continuum of semantically differently strong connected elements. The borders between "free" and "firm" can not be specified clearly. On the other hand, the main problem of the statistical approach is that the frequency and semantic firmness of word combinations do not correlate directly. Not all high frequent word combinations are also firm. One finds typical collocations in all ranks of the frequency distribution [15, 16].

In our lexical database, the (meaningful) collocations are manually selected from the first 500 occurrences of each grammatical structure listed by The Sketch Engine and cross-checked against the Slovak National Corpus concordances.

The statistical results vary, they depend both on the used statistical method and the quality and accuracy of taggers and lemmatisers, the precision rates whereof are different. It means that we have to compare very long lists of indexes from different scores. The table 1 shows the identity of collocation candidates between scores within the first 50 rows.

	T-score	MI	MI3	log likelihood	min. sensitivity	salience
Abs. freq	73.9	5.4	5.4	54.9	34.4	36.6
T-score		5.0	53.6	70.1	62.0	44.4
MI			32.7	24.5	14.0	36.0
MI3				75.0	57.1	81.6
log likelihood					71.4	69.4
min. sensitivity						60.0

Table 1. Comparing collocations – identity between different measures, in percents.

3 Technical implementation of the lexical database

Since the dictionary has been conceived from the beginning as a collaborative project involving several contributors, the choice of the working environment has been driven by several requirements – easy remote editing, access control list, revision history, communication between editors. These requirements can be easily met by deploying a wiki based software, we have chosen MediaWiki software system, with MySQL as a relational database backend.

MediaWiki is written in the PHP programming language and has many attractive options for the intended purposes, among them the possibility to use templates (a kind of macro) for better handling of repeating text parts. Templates are basically predefined text snippets in wiki-format with additional specialized markup for accommodating passing of arguments which are dynamically loaded inside another page. More on this in section 6.2.

While a wiki system has proved as highly suitable for the task of creating the dictionary, the way of representing the dictionary information to the end user is still an open question, the layout provided by the wiki-entries being probably not the most appealing and useful one.

lemma	MI-score	lemma	T-score
anakreontov	13.03	ľudový	50.70
hiawathowa	13.03	text	23.91
kancionálový	12.99	táto	20.87
paraliturgický	12.96	tá	19.91
brelov	12.77	duchovný	19.54
rózsov	12.77	spievať	19.41
švihrovsky	12.77	nový	19.29
hymnický	12.74	**labutí**	17.57
dylanový	12.55	populárny	16.65
pestúnkina	12.31	vianočný	15.85
slávikoví	12.19	pieseň	15.55
labutí	12.07	známy	15.07
schubertový	12.03	zaspievať	14.04
legendický	12.01	**nábožný**	13.63
švihrovský	11.90	titulný	13.17
pijácky	11.86	náboženský	12.78
podkovitý	11.77	**hymnický**	12.61
cherubínsky	11.77	obľúbený	12.61
povaľačský	11.77	ľúbostný	11.54
regrútsky	11.70	rómsky	11.14
symfonia	11.36	večný	10.78
lennonový	11.36	smutný	9.04
silvánov	11.27	mariánsky	8.51
mický	11.12	svadobný	8.49
zaspievanie	11.09	oslavný	7.99
carlina	10.96	rusínsky	7.99
barnabášov	10.96	tanečný	7.90
kramársky	10.82	skladať	7.83
trampský	10.79	lyrický	7.11
nábožný	10.78	pohrebný	7.10

Table 2. Differences in the lists of collocation candidates extracted by MI-score and T-score, lemma *pieseň*. Words in boldface are shared between top 30 occurrences of both scores.

4 Prerequisities

In the initial phase of the project, the collocations were obtained from Slovak National Corpus (SNK), version *prim-3.0* containing about 330 million tokens. Halfway during the work on the database, a new version of the SNK has been released (*prim-4.0*), bringing the number of tokens up to 530 million, which faced us with a dilemma: as the new version had not only substantially increased in the volume, but also improved lemmatization and morphology annotation, it would be advantageous to use this new information, but on the other hand, changing the input data would require to go through and redo all the entries already done. At the end, we decided to use the new version for new entries and analyse the collocational profiles with respect to changed statistical measures in order to evaluate the changes brought by a new corpus.

There is also the question of which variant of the corpus to use – there are three main flavours of the Slovak National Corpus database, *prim-4.0-public-all* contains all the texts, *prim-4.0-public-sane* contains only texts that satisfy certain requirements (correct diacritics, non linguistic texts, no texts written by Slovak minorities outside of Slovakia proper, some controversial writers removed), *prim-4.0-public-vyv* is a balanced corpus, containing 1/3 newspaper texts, 1/3 fiction, 1/3 scientific texts (see Tab. 3 for comparison).

Version	number of tokens		
	-all	-sane	-vyv
prim-3.0-public-	339 063 215	319 644 966	199 822 572
prim-4.0-public-	526 082 640	507 101 251	254 236 903

Table 3. Comparing versions 3.0 and 4.0 of the Slovak National Corpus.

Corpus	prim-4.0-public-all	prim-4.0-public-sane	prim-4.0-public-vyv
Identity: sane+vyv		75.5%	
Identity: all+sane	93.5%		
Identity: all+vyv			74.5%
Identity: all+sane+vyv	73.4%		
Isolated occurrences	9.7%		

Table 4. Comparing identity of collocation candidates of the word *pieseň* (song) in three different versions of the Slovak National Corpus, version 4.0.

5 Basic structure of the database

The database serves two different purposes – the first is to build a Slovak language collocation dictionary, the second one to build a (semi)bilingual dictionary of German collocations with Slovak equivalents [18, 20]. These two projects share the same database and the same MediaWiki instance, and (to an extent) use the same methods and guidelines regarding the collocation profiles. However, logically these are two separate projects. In this paper we deal exclusively with the Slovak dictionary.

The database macrostructure is simple – all the entries are equal, each entry corresponds to one MediaWiki page, we are using neither subpages[5] nor redirects[6]. A page is named by an entry lemma, in case of clash between German and Slovak (e.g. Internet, System), the Slovak page adds the string '(sk)' to the page name, so that the pages will be named 'Internet (sk)', 'System (sk)'. Unfortunately, MediaWiki automatically converts the names to titlecase, otherwise the compulsory capitalization of German nouns would distinguish between German and Slovak entries. Slovak lexical entries are differentiated from other pages (system pages, German entries, user discussions) by the category they belong to (one of Slovak Nouns, Slovak Adjectives, Slovak Verbs, Slovak Particles).

6 Structure of an entry

An entry page consists of three main sections: *Významy* (Meanings), *Kolokácie* (Collocations), *Externé odkazy* (External links). While the structure of *Významy* and *Externé odkazy* is the same for all the parts of speech and these sections do not have any substructure, the structure of *Kolokácie*, the most important section, is more complicated [19].

6.1 Významy

This section ("meanings") contains a bullet list of descriptions of different definitions of the lexeme. We do not split the collocations according to polysemy (or homonymy) of the base noun inside one part of speech category at all, neither we distinguish between homonyms in collocations. This was a deliberate design decision, based on two observations: First, often a collocation is not clearly attributable to a specific meaning; let alone trying to define and distinguish meanings, which is traditionally a very cumbersome task, where no general consent could be achieved. This was not seen as a task for this project and would unnecessarily considerably slow down the dictionary constructions and open door to endless discussions inside and outside the project team about the distinction of individual meanings.

6.2 Kolokácie

All the collocation data are contained in this section. The detailed structure is differentiated according to part of speech the entry stands for. For nouns, it is divided into two subsections for the singular and plural, reflecting the fact that collocates often exhibit different phenomena according to the grammatical number of the base noun. Each of these subsections is further divided into many subsubsections, each for a specific collocation combination (see Fig. 1, 2).

The subsubsections' naming scheme encodes some human readable information about the collocations, with the base noun marked by the string *Sub1Xxx*, where *Xxx* is the abbreviation of the noun's case (so the whole string will be one of *Sub1Nom, Sub1Gen, Sub1Dat, Sub1Aku, Sub1Lok, Sub1Ins*). We are ignoring the Slovak vocative controversy by conflating (semantic) vocatives with the nominative case – fortunately, it just happened that none of the nouns chosen for the collocation dictionary is from the set of those few Slovak words that have a morphological vocative (either having retained their Old Slavic vocative forms[7], or developing a 'new vocative', common for some proper nouns and family relationships[8].

The other part of the subsubsection name reflects describes the neighbouring word part of speech, so it can be one of *Sub2, Verb, Atr* (another noun, verb, attribute). *Atr* subsumes adjectives,

[5] A subpage is a page that is subordinate to its parent page. Subpages can be used to implement a whole hierarchy (tree structure) of entries, which – considering lexicographic use – can be used to distinguish between homonymy and polysemy [9]. However, given the small number of entries, we decided to refrain from this.

[6] A redirect is a page that has no data by itself; it just refers to another page.

[7] e.g. *otče, pane, bože*

[8] e.g. *babi, mami, Zuzi, Feri*

pronouns, particles or numerals. This string is positioned either to the left or to the right of the previous base noun string, depending on the predominant position of the word in collocations (but including also the collocations with a different word order). The strings are concatenated with a plus sign, so e.g. the whole subsubsection name *Verb + Sub1Aku* indicates that the subsubsection contains collocation of verb and base noun in acusative (not necessarily in this order).

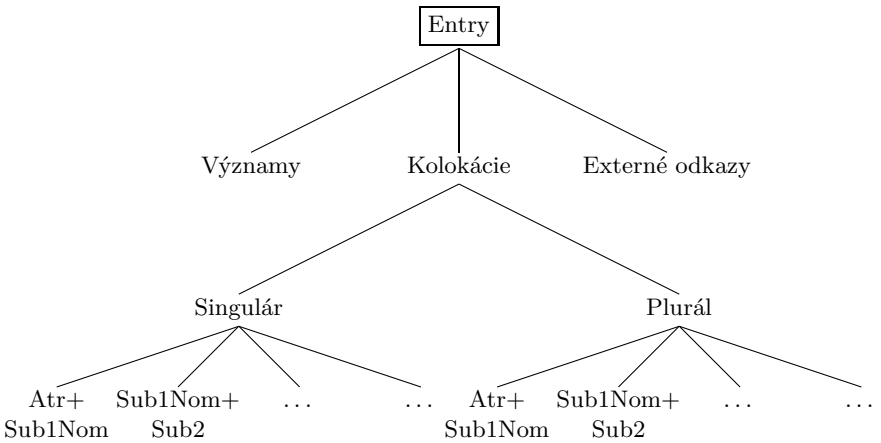

Fig. 1. Entry structure diagram for nouns

Sub1	Sub2	Verb	Atr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Atr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Atr
...	...	...	...
Sg Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Atr
Pl Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Atr
Pl Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Atr
...	...	...	...
Pl Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Atr

Fig. 2. Matrix for the entry structure of a noun

6.3 Externé odkazy

This section is populated by several macros (templates), providing links to external resources. Each macro has one parameter, equal to the identification of given word in the target database – mostly the same as the lemma, different only in case of homonyms (differentiated at the target). The macros construct an URL pointing to an external resource and insert it as an http hyperlink into the rendered page. The macros in use are `{{ma|...}}` to link to morphologic database (this macro is intended to record relations between full word paradigms and the collocation dictionary entries, both for the end user and for eventual computer processing), `{{slovnik|...}}` to link to dictionaries[7] published at the Ľ. Štúr of Linguistics WWW page, `{{linky|...}}` to point to several search engines, such as Google[1], Ask[2], Yahoo[3], Cuil[4], as well as the Slovak National Corpus[6]. The latter two templates are meant for human consumption, not for computer parsing (due to somewhat unpredictable nature of the target data). In case we need to either add or remove an external data source (e.g. a search engine), or if the form of URL parameters changes,

we need to modify just the template, and the change will be automatically reflected across all the database entries.

7 Automated database processing

There are several options for automated data modification. First and most obvious is to access the SQL backend directly, reading and modifying the tables. However, this method requires detailed knowledge of internal MediaWiki database structure, and modifying would have to be done with a great care, in order not to disrupt the database and introduce structural inconsistencies.

Much better way is to use a MediaWiki API, designed for a remote access. As the MediaWiki is probably the most widely used Wiki framework, there is a plethora of tools available[5] for automated processing in various programming languages. However, we settled on using a slightly different approach – WikipediaFS[8], a fuse-based[12] filesystem that presents remote WikiMedia installation as a fake filesystem, so that the pages can be read and written as simple text files, either for automated scripted processing or to be edited with an ordinary text editor. The advantage of WikipediaFS over using MediaWiki API is the availability of plain text, filesystem like view of the data, which makes it easy to use standard UNIX command line tools for text processing (`sed, awk, grep, ...`). We used WikipediaFS and some simple scripts to add automatically the abovementioned links to external resources to all the entries in the database.

8 Collocation entry microlanguage

The lexical database has been designed with a goal of a human readable collocation dictionary in mind, published both online and in printed form. However, the importance of the need to keep the data in computer readable format cannot be stressed enough – if nothing else, to automatise the typographic formatting process for the printed version, and indexing for the online version. Therefore the entry microformat is designed to be computer readable, except of some minor exceptions, where the (complete) readability stands in the way of human interaction.

Each collocation can be though of as consisting of two units: the base noun and the collocate. The collocates are normalised (lemmatised), and the collocation is written with the base in its corresponding case/number. The exception is only for the combination *Atr + Sub1Nom*, which is so frequent that we omit the base in nominative, if it follows the attribute. Auxiliary particles/pronouns are sometimes rearranged, to fit the syntactical requirements of the base (this applies mainly to the reflective pronouns *sa, si* in combination with infinitives). From this follows that the parser must include the morphology generator in order to recognise the base noun in other forms than nominative singular, and a complete automatised parsing is difficult without including some sort of syntactical rules into the parser. Collocate is terminated by the | (`U+007C VERTICAL LINE`) character surrounded by whitespace. The vertical line has to terminate also the ultimate collocate in the subsubsection. If there are no collocates for a given collocation pattern, the entry consists of a single vertical line character in a separate line. Optional words (which are sometimes present in a given collocation) are enclosed in parentheses, separated by the rest of collocation by a whitespace or punctuation. Parentheses adjoined to a word specify optional prefixes or suffixes (mostly verb negation or aspect modifier). Variants in words (two or more words that do not change the collocation meaning and are approximately equally frequent) are separated by a slash, three dots (ellipsis, ...) denote incomplete variant enumeration (signalling that there are more variants occurring in the corpus than given, usually these variant components belong to a specific lexico-semantic group). Special indefinite pronouns (*niekto, niečo, ...*) serve as wildcard valency markers which stand for a general class of aminate/inanimate nouns (and thus signal that the collocation is too broad to be automatically parsed).

There are on average 173 collocations per entry – the distribution of entry sizes is depicted on Fig. 4. We see that the symmetry is slightly skewed in favour of small number of bigger sized entries (the median is 157). The entry with fewest number of collocations is *kára* (cart, barrow), with 40 collocations, the highest number has the word *svet* (world) – 584 collocations. However, we have

to realise that the exact number of collocations per entry is subject to several arbitrary conditions, among them the level of detail in describing collocation variants, inclusion of otherwise optional ellipsis and indefinite pronouns, and in general subjective evaluation of collocation candidates by a lexicographer compiling the entry.

```
==Atr + Sub1Gen==

neznalý pomerov | z chudobných pomerov | znalý pomerov |

==Sub2 + Sub1Gen==

demokratizácia pomerov | konsolidácia pomerov | kritika pomerov |
neznalosť pomerov | obraz (politických / reálnych / ... ) pomerov |
stabilizácia pomerov | úprava pomerov | usporiadanie pomerov |
zlepšenie pomerov | zmena spoločenských / vlastníckych pomerov |
znalec (našich domácich) pomerov | znalosť tunajších pomerov |

==Verb + Sub1Gen==

pochádzať z (dosť) chudobných / skromných pomerov |
```

Fig. 3. Fragment of a collocation entry, word *pomer*

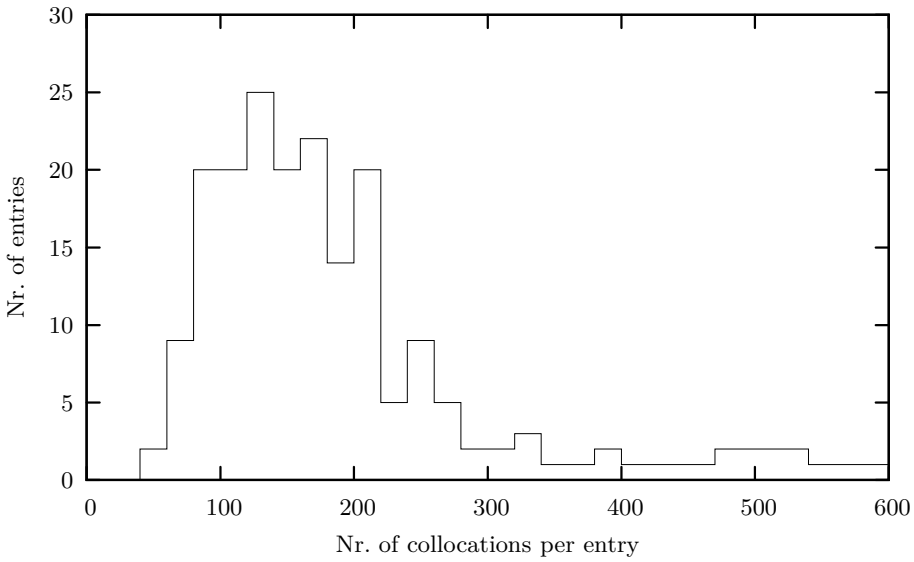

Fig. 4. Distribution of number of collocations per noun, bin size = 20

9 Further plans & Conclusion

The plan for the first phase of the project is to create a dictionary of noun collocations, with the number of entries exceeding 500. Currently, the database contains collocation profiles of 190 nouns and 38 particles.

After the first phase, a new methodology for a dictionary of other parts of speech will be delineated and the dictionary will be extended. It is expected that by that time a new version of the Slovak National Corpus database will be available, and already existing entries could be cross validated against these new data. The dictionary will be a valuable contribution to modern Slovak language lexicography, reflecting real language usage by being based on the real data from the Slovak National corpus.

From the theoretical point of view, research of collocations will add to our knowledge about the collocability of words, presented collocation database can serve as a base for confrontational Slovak language research. Collocations per se form an inseparable part of many different kinds of dictionaries, and they are especially important in language teaching, giving examples of real language usage. We believe that the collocation dictionary will be used in teaching Slovak as a foreign language, since the mastery of idioms is a sign of a true language competency.

References

[1] http://www.google.com.
[2] http://www.ask.com.
[3] http://www.yahoo.com.
[4] http://www.cuil.com.
[5] Botwiki. http://botwiki.sno.cc/wiki/Manual:Frameworks. A wiki for documenting and testing bots. Retrieved 2009-06-08.
[6] Slovak National Corpus. http://korpus.juls.savba.sk.
[7] Slovenské slovníky. http://slovnik.juls.savba.sk.
[8] Blondel, M. WikipediaFS. http://wikipediafs.sourceforge.net/. Retrieved 2009-06-08.
[9] Garabík, R. & Špirudová, J. (2009). Design of a New Slovak-Czech Lexical Database. In Garabík, R. (Ed.), *Metalanguage and Encoding Scheme Design for Digital Lexicography*, (pp. 71–76)., Bratislava, Slovakia. Tribun.
[10] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The sketch engine. *Information Technology, 105.*
[11] Majchráková, D. & Ďurčo, P. (2009). Compiling the First Electronic Dictionary of Slovak Collocations. To be published.
[12] Szeredi, M. Filesystem in Userspace. http://fuse.sourceforge.net/. Retrieved 2009-06-08.
[13] Tvrdý, P. (1931). *Slovenský frazeologický slovník.* Trnava: Spolok sv. Vojtecha.
[14] Tvrdý, P. (1933). *Slovenský frazeologický slovník. Druhé doplnené vydanie.* Praha and Prešov: Nákladom Československej grafickej unie, úč. spol.
[15] Čermák, F. (2006a). Kolokace v lingvistice. In Čermák, F. & Šulc, M. (Eds.), *Kolokace. Studie z korpusové lingvistiky. Sv. 2.*, (pp. 9–16)., Prague, Czech Republic. Nakladatelství Lidové noviny, Ústav Českého národního korpusu.
[16] Čermák, F. (2006b). Statistical Methods for Searching Idioms in Text Corpora. In Häcki-Buhofer, A. & Burger, H. (Eds.), *Phraseology in Motion 1. Methoden und Kritik*, (pp. 33–42). Baltmannsweiler (Schneider Verlag Hohengehren).
[17] Ďurčo, P. (2007a). Collocations in Slovak (Based on the Slovak National Corpus). In Garabík, R. & Levická, J. (Eds.), *Computer Treatment of Slavic and East European Languages*, (pp. 43–50)., Bratislava, Slovakia. Tribun.
[18] Ďurčo, P. (2007b). O projekte nemecko-slovenského slovníka kolokácií. In Baláková, D. & Ďurčo, P. (Eds.), *Frazeologické štúdie V. Princípy lingvistickej analýzy vo frazeológii*, (pp. 70–93)., Ružomberok, Slovakia. Katolícka univerzita v Ružomberku.
[19] Ďurčo, P. (2007c). Zásady spracovania slovníka kolokácií slovenského jazyka. http://www.vronk.net/wicol/images/Zasady.pdf. Online documentation.
[20] Ďurčo, P. (2008). Zum Konzept eines zweisprachigen Kollokationswörterbuchs. Prinzipien der Erstellung am Beispiel Deutsch-Slowakisch. In Hausmann, F. J. (Ed.), *Collocations in European lexicography and dictionary research. Lexicographica*, volume 24, (pp. 69–89)., Tübingen, Germany. Max Niemeyer Verlag.