Contrastive Dictionary of German and 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. In the article we discuss ongoing work concerning a confrontational German-Slovak collocation lexical database. The database consists of two parts, a section of German collocations with Slovak equivalents and a section of Slovak collocations. Intended size of the database is several hundred words of different parts of speech (nouns in the first phase of the project) for each of the languages, together with their collocation profiles. The database uses MediaWiki engine and a wiki-based approach to article editing and collaborative work of a team of lexicographers.

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) [21, 22]. 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 German and Slovak collocations is being compiled at the University of St. Cyril and Methodius, Trnava in cooperation

* 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. Parts of this article have been published in [27].

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 [23]. Description models on the basis of collocational matrices are elaborated also for verbal, adjectival, adverbial and partical collocations.

We exclude regular systematic terminological and proprial collocations from the database, leaving only irregular systematic collocations (idioms, phrasemes), regular text-collocations (e.g. *zimná rekreácia*) and fixed text-systematic collocations (e.g. *krájať nadrobno, hovoriť úsečne*).

2 Obtaining collocation profiles

To obtain Slovak collocation profiles from a large lemmatized corpus we are using the sketch engine[4] [18] – a corpus tool which generates word sketches, i. e. corpus based summaries of grammatical and collocational behaviour of a word. Disadvantages of the 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 [19].

In our lexical database, the Slovak 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 German collocations are obtained from the IdS corpus [7], DWDS corpus [6] and the Wortschatz-Portal [13]. We use the most frequent words from [16], updated by the data from [17] (currently the only one contemporary frequency list of German words, created out of a specialized corpus reference texts). Unfortunately, there is no word sketch created for German language, however we are working on a preliminary version. We then add Slovak equivalents to the German collocation in the database.

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

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.

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 7.2.

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.

4 Building Slovak collocations

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.

5 Slovak equivalents of German collocations

German section of the database consists of German language collocations with their equivalents in Slovak. During the construction of the database, we observed several common patterns in the equivalency[14]:

Monosemic German words are reflected in monosemic equivalency – in the whole collocation profile, there is consistently one Slovak equivalent. Examples of such words are Schutz – ochrana (protection), Reise – cesta (journey); Schüler – žiak (pupil); Sommer – leto (summer).

effektiver Schutz	efektívna *ochrana*
Schutz gegen Inflation	*ochrana* proti inflácii
jmdm. Schutz gewähren	poskytnúť niekomu *ochranu*

Table 1. Example Slovak collocation equivalents for a monosemic German word *Schutz*

Polysemic German words cover several meanings and consequently they are assigned several Slovak equivalents – sometimes even one German meaning is translated into several (closely related) Slovak words.

einfacher Satz	holá *veta*
entscheidender Satz	1. rozhodujúca *veta*, 2. rozhodujúci *set*
ermäßigter Satz	znížená *sadzba*
der zweite Satz des (Volleyball) spiels	druhý *set* hry (volejbalu)
Satz Autoreifen	*sada* pneumatík
Satz des Kaffees	*usadenina* z kávy
Satz Fische	*násada* rýb
Satz von Anordnungen	*súbor* predpisov
mit einem jähen / schnellen Satz an der Tür sein	rýchlym *skokom* byť pri dverách

Table 2. Example Slovak collocation equivalents for a polysemic German word *Satz*

Monosemic foreign German words (*Fremdwörter, Lehnwörter*) are usually reflected by monosemic Slovak equivalents, thanks to the common Greco-Latin heritage often of the same etymology. Examples: Reform – reforma (reform); Symbol – symbol (symbol); Student – študent (student).

Polysemic foreign German words cover several meanings, and are usually assigned several Slovak equivalents, thanks to etymological divergence there are often several possible Slovak words (usually a loanword and its Slovak equivalent). Examples: Reaktion – reakcia, odozva, odpoveď, chemická reakcia, protipôsobenie (reaction); Situation – situácia, stav, pomery (situation); Transport – transport, preprava, doprava (transport); Qualifikation – kvalifikácia, posúdenie (qualification).

Polyequivalence – there are words, whose equivalents are not just a single translation, but a set of several synonyms that are sometimes freely interchangeable, but sometimes not. We assign the Slovak equivalents according to their typical usage in the Slovak language. E.g., German word *Sinn* (sense, meaning, point, idea, consciousness, feeling,. . .) can be translated by *význam, zmysel, cit, pochopenie, myseľ*, but the collocations of the equivalents are somewhat rigid and not all of them are interchangeable.

im engeren Sinn des / eines Wortes	v užšom *zmysle* slova
keinen Sinn in etwas sehen	v niečom nevidieť žiadny *význam* / *zmysel*
jmds. Sinn ist nur auf dieses eine Ziel gerichtet	niekoho *myseľ* je napriamená len na jeden cieľ
frohen Sinnes sein	byť veselej *mysle*

Table 3. Example of Slovak collocation equivalents diversity for the German word *Sinn*

The same applies for idiomatic word usage. E.g. the German word *Stunde* (hour) has more or less straightforward meaning, translated by Slovak *hodina*, but when the German collocations cover the idiomatic usage, we have to use corresponding Slovak idioms.

in einer schwachen Stunde	v slabej *chvíli, chvíľke* - *v slabej *hodine*
die richtige Stunde abwarten	vyčkať správny *okamih*
in einer stillen Stunde	vo chvíľke *pokoja*
zur richtigen Stunde	v pravý *čas*

Table 4. Example of idiomatic usage of the German word *Stunde*

6 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 [24, 26]. These two projects share the same database and the same MediaWiki installation, and (to an extent) use the same methods and guidelines regarding the collocation profiles. The databases are distinguished on the logical level, by marking each entry as belonging to one of the Slovak part of speech collocation categories `Slovak Nouns`, `Slovak Adjectives`, `Slovak Verbs`, `Slovak Particles` or to the `German collocation` category. The rest of the pages (not belonging to either category) are system, user or administrative pages, or user discussions.

The database macrostructure is simple – all the entries are equal, each entry corresponds to one MediaWiki page, we are using neither subpages nor redirects. 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 could be used to distinguish between German and Slovak entries.

7 Structure of an entry

Overall structure of an entry is identical for both German and Slovak parts of the database. They differ in the language of section titles, where we use German terms for the German entries and Slovak terms for Slovak entries. In the following text, we describe both version together, putting German section names first, followed by Slovak ones.

An entry page consists of three main sections: *Bedeutung* or *Významy* (Meanings), *Kollokationen* or *Kolokácie* (Collocations), *Links* or *Externé odkazy* (External links). While the structure of *Bedeutung* and *Links*, or *Významy* and *Externé odkazy* is the same for all the parts of speech and these sections do not have any substructure, the structure of *Kollokationen* or *Kolokácie*, the most important section, is more complicated [25].

7.1 Bedeutung, 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; second, trying to define and distinguish meanings is traditionally a very cumbersome process, 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.

7.2 Kollokationen, 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, 3, 4).

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, Sub1Akku* or *Sub1Aku, Sub1Lok, Sub1Ins*). We are ignoring the Slovak vocative controversy by conflating (semantic) vocatives with the nominative case – fortunately, none of the nouns chosen for the collocation dictionary is from the set of those few Slovak words that have a morphological vocative.

The other part of the subsubsection name reflects describes the neighbouring word part of speech, so it can be one of *Sub2, Verb, Attr* or *Atr* (another noun, verb, attribute). *Attr* or *Atr* subsumes adjectives, pronouns, particles and 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 + Sub1Gen* indicates that the subsubsection contains collocation of verb and base noun in genitive (not necessarily in this order).

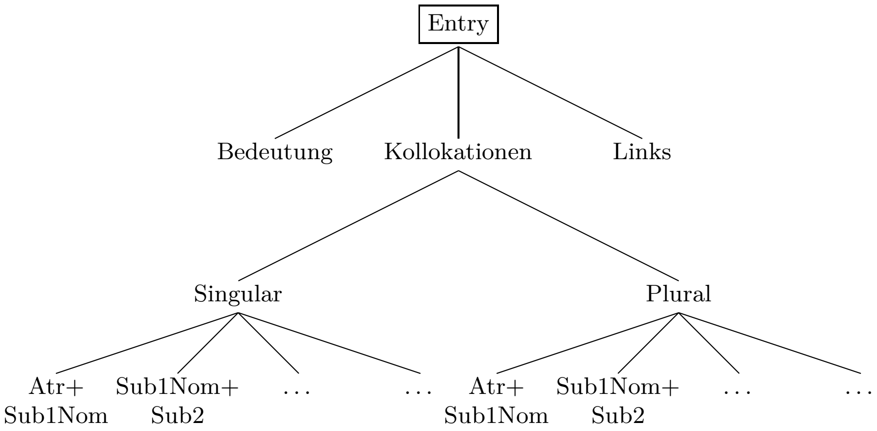

Fig. 1. Entry structure diagram for German nouns

7.3 Links, 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.

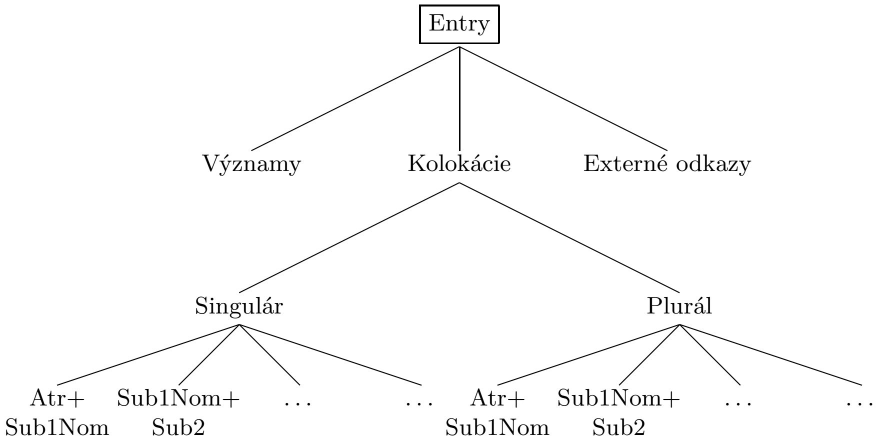

Fig. 2. Entry structure diagram for Slovak nouns

Sub1	Sub2	Verb	Atrr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Sg Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Sg Akku	Sub1Akku+Sub2	Sub1Akku+Verb	Sub1Akku+Attr
Pl Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Attr
Pl Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Attr
Pl Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Attr
Pl Akku	Sub1Akku+Sub2	Sub1Akku+Verb	Sub1Akku+Attr

Fig. 3. Matrix for the entry structure of a German noun

Sub1	Sub2	Verb	Atr
Sg Nom	Sub1Nom+Sub2	Sub1Nom+Verb	Sub1Nom+Atr
Sg Gen	Sub1Gen+Sub2	Sub1Gen+Verb	Sub1Gen+Atr
Sg Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Atr
Sg Aku	Sub1Aku+Sub2	Sub1Aku+Verb	Sub1Aku+Atr
Sg Lok	Sub1Lok+Sub2	Sub1Lok+Verb	Sub1Lok+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 Dat	Sub1Dat+Sub2	Sub1Dat+Verb	Sub1Dat+Atr
Pl Aku	Sub1Aku+Sub2	Sub1Aku+Verb	Sub1Aku+Atr
Pl Lok	Sub1Lok+Sub2	Sub1Lok+Verb	Sub1Lok+Atr
Pl Ins	Sub1Ins+Sub2	Sub1Ins+Verb	Sub1Ins+Atr

Fig. 4. Matrix for the entry structure of a Slovak noun

Slovak language macros. 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[11] 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[10]. 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.

German language macros. In the German section, the entries use only a single macro to link to all of the external sources – `{{links-de|...}}` links to several German online dictionaries: dict.cc German-English dictionary [9] (includes full morphology paradigms), the LEO German-English dictionary [8], DWDS monolingual dictionary [6] and Zoznam German-Slovak dictionary [12].

8 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[15], a fuse-based[20] 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.

9 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 collocation is written with the base in its corresponding case/number, there is only one exception – in the Slovak database, the combination *Atr + Sub1Nom* is so frequent that we decided to omit the base if in nominative, when it immediately 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). German reflexive verbs that take the subject in dative are marked with a special qualifier `(Dat.)`, e.g. `sich (Dat.) die Augen reiben; sich (Dat.) die Augen verderben; sich (Dat.) die Augen wischen`.

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.

In the Slovak database, each 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.

In the German database, collocations in each subsection are organized in a two column table, with German collocates on the left and Slovak equivalents on the right. We are using standard MediaWiki table syntax, starting each table with a following code snippet (preamble and table header):

```
{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
```

Each collocation is in one table row, the rows are separated by a `|-` string (a vertical line followed by a hyphen-minus), in each row there is a German collocate, followed by a string `||` (two vertical lines), followed by a Slovak collocate.

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).

In the Slovak section, there are on average 173 collocations per entry – the distribution of entry sizes is depicted on Fig. 7. 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.

In the German section, the average number of collocations per entry is 195.5 (see Fig. 8), the median is 150. The entry with the fewest number of collocations is *September*, with 21 collocations, the entry with the highest number is *Kind* (child), with 703 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. The subjective differences are even more pronounced when comparing two different languages, where also the language competence of the lexicographer plays its role (if they are not native speakers of the language), and also different methodology of obtaining collocation profiles.

```
==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. 5. Fragment of a Slovak collocation entry, word *pomer*

10 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 Slovak database contains collocation profiles of 190 nouns and 38 particles, the German database contains 280 collocation profiles, all of them are nouns.

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 Slovak language entries could be cross validated against these new data. The dictionary will be a valuable contribution to modern

```
===Attr + Sub1Nom===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|beigegebene Abbildungen || priložené obrázky
|-
|unzüchtige Abbildungen || nemravné / oplzlé obrázky
|-
|verschiedene Abbildungen || rôzne obrázky
|-
|zahlreiche Abbildungen || početné obrázky
|}

===Sub1Nom + Sub2===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|Abbildungen aller Art || rôznorodé / rozmanité zobrazenia
|-
|Abbildungen durch Linsen || zobrazenia prostredníctvom šošovky
|-
|Abbildungen in Band xy || obrázky vo zväzku xy
|-
|Abbildungen im Text || obrázky v texte
|-
|Abbildungen oder Darstellungen || zobrazenia alebo znázornenia
|-
|Abbildungen von Funden / Gegenständen / ... || obrázky nálezov / predmetov / ...
|-
|Beschreibungen der Abbildungen || popisy obrázkov
|-
|Zeichnungen oder Abbildungen || kresby alebo obrázky
|}

===Sub1Nom + Verb===

{|class="wikitable"
|-
!Deutsch - Nemecky
!Slowakisch - Slovensky
|-
|Abbildungen enthalten etwas || obrázky obsahujú niečo
|-
|Abbildungen geben Vorstellung / Zeugnis / ... || obrázky poskytujú
                                                  predstavu / svedectvo / ...
|-
|Abbildungen veranschaulichen etwas || obrázky znázorňujú niečo
|-
|die Abbildungen zeigen jn, etwas || obrázky ukazujú niekoho, niečo
|}
```

Fig. 6. Fragment of a German collocation entry, word *Abbildung*

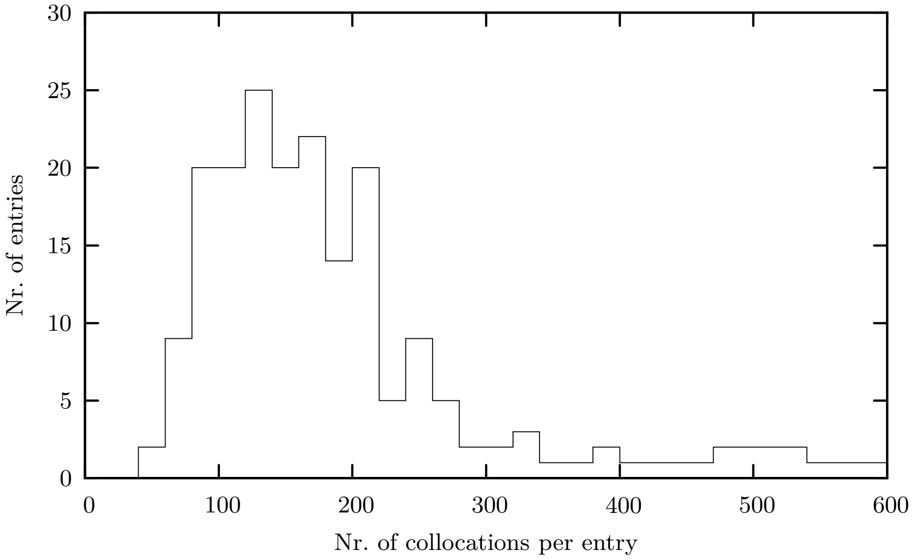

Fig. 7. Distribution of number of collocations per noun in the Slovak section, bin size = 20

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.

The bilingual German-Slovak collocation database will offer excellent possibilities for contrastive linguistic studies, and will be similarly useful for Slovak speakers in learning German as a foreign language as well as for German speakers in learning Slovak.

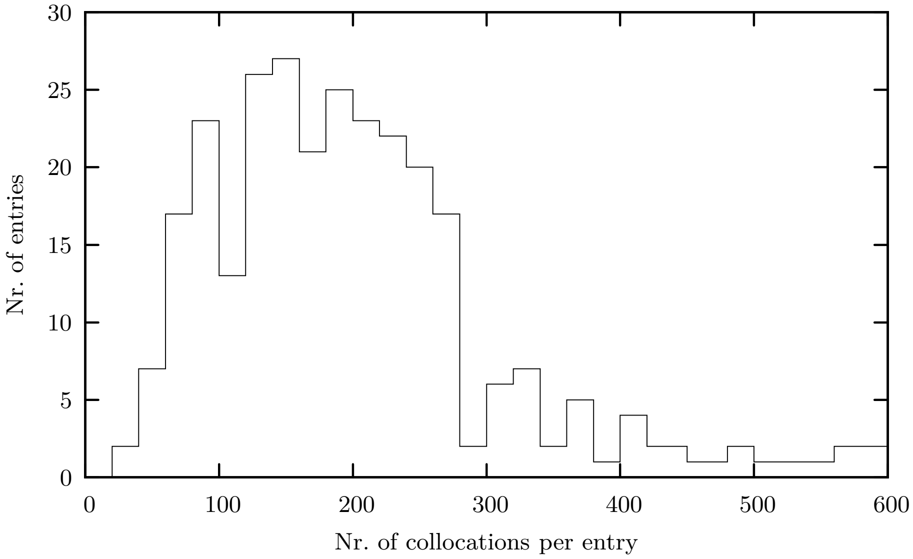

Fig. 8. Distribution of number of collocations per noun in the German section, bin size = 20

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] Das Digitale Wörterbuch der deutschen Sprache des 20. Jh. http://www.dwds.de.

[7] Das Portal für die Korpusrecherche in den Textkorpora des Instituts für Deutsche Sprache. http://www.ids-mannheim.de/cosmas2/.

[8] Deutsch-Englisch Wörterbuch, Ein Online-Service der LEO GmbH. http://dict.leo.org.

[9] dict.cc, English-German Dictionary. http://dict.cc.

[10] Slovak National Corpus. http://korpus.juls.savba.sk.

[11] Slovenské slovníky. http://slovnik.juls.savba.sk.

[12] Web slovník. http://webslovnik.zoznam.sk.

[13] Wortschatz Universität Leipzig. http://www.ids-mannheim.de/cosmas2/.

[14] Banášová, M. (2008). Polysemie und Polyäquivalenz der Kollokationen im Deutsch-slowakischen Kollokationswörterbuch. In Ďurčo, P. (Ed.), *5. Kolloquium zur Lexikographie und Wörterbuchforschung. The Fifth International Colloquium on Lexicography/Feste Wortverbindungen und Lexikographie/Fixed word combinations and Lexicography.*, Bratislava, Slovakia. (in press).

[15] Blondel, M. WikipediaFS. http://wikipediafs.sourceforge.net/. Retrieved 2009-06-08.

[16] Glaboniat, M., Muller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch. Gemeinsamer Europäischer Referenzrahmen.* Berlin, Germany.: Langenscheidt.

[17] Jones, R. L. & Tschirner, E. (2005). *A Frequency Dictionary Of German.* Routledge.

[18] Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The sketch engine. *Information Technology*, *105.*

[19] Majchráková, D. & Ďurčo, P. (2009). Compiling the First Electronic Dictionary of Slovak Collocations. To be published.

[20] Szeredi, M. Filesystem in Userspace. http://fuse.sourceforge.net/. Retrieved 2009-06-08.

[21] Tvrdý, P. (1931). *Slovenský frazeologický slovník.* Trnava: Spolok sv. Vojtecha.

[22] Tvrdý, P. (1933). *Slovenský frazeologický slovník. Druhé doplnené vydanie.* Praha and Prešov: Nákladom Československej grafickej unie, úč. spol.

[23] 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.

[24] 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.

[25] Ďurčo, P. (2007c). Zásady spracovania slovníka kolokácií slovenského jazyka. `http://www.vronk.net/wicol/images/Zasady.pdf`. Online documentation.

[26] 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.

[27] Ďurčo, P., Garabík, R., Daniela, M., & Ďurčo, M. (2009). Dictionary of Slovak Collocations. In *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography.*, (pp. 128–137)., Warsaw, Poland. Institute of Slavic Studies, Polish Academy of Sciences.