PRACE FILOLOGICZNE

Warszawa 2012 TOM LXIII

RADOVAN GARABÍK, MÁRIA ŠIMKOVÁ
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences, Bratislava

THE SLOVAK NATIONAL CORPUS AND ITS CORPUS LINGUISTIC RESOURCES

SŁOWA KLUCZOWE: Narodowy Korpus Języka Słowackiego, anotacja, narzędzia korpusowe, kolokacje, korpus języka mówionego

KEYWORDS: Slovak National Corpus, annotation, corpus tools, collocations, spoken language corpus

1. Slovak National Corpus

The history of the Slovak lexical database and corpus is only 20 years old. There were theoretical works on the integral concept of the future corpus of the Slovak language and lexical database (namely by J. Horecký, A. Jarošová), a theoretical computer model of the Slovak language (E. Páleš, 1994), and the main work was a practical collection of texts in electronic form and their first linguistic analyses. This corpus of texts of the Slovak language was gradually made available from 1996 up to 2002 for internal use within the Ľ. Štúr Linguistic Institute of the Slovak Academy of Sciences. In its final phase, the 30-million-word corpus included mainly journalistic texts, various texts from professional proceedings and journals, and a small quantity of fiction. A specific part of the corpus consisted of electronic versions of the lexicographic productions of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences. The first corpus / lexical database of the Slovak language was built solely for the purpose of creating a new dictionary of the Slovak language.

Moreover, user demand for publicly accessible linguistic information began to increase throughout the late 1990s. The lexicographers required an increase in the volume and improvement of the structure of corpus texts to increase their work efficiency. More demands emerged within the context of Slovakia's expected acceptance into the European Union. After consideration was given to the optimal place and method for the systematic building of a new corpus with internationally comparable parameters, the current project was developed. Preparatory work was launched after the project had been approved by the government of the Slovak Republic in 2002. The team comprises eight members and dozens of other collaborators.

Currently, the Slovak National Corpus (SNK) is a large, representative corpus of modern written Slovak (since the 1953 orthography reform). Access to the corpus query interface is publicly available after a simple registration process for educational and other non-commercial research purposes. The rules to create a representative/balanced sample of written texts of the current Slovak language were only approximate: one third will consist of journalistic texts, the second of fictional texts and the final third of specialised and non-fiction texts. Translations are prominent in the two latter groups, as they have a special position in small national and language societies (such as Slovakia). Moreover, they were very poorly represented in the previous lexicographical manuals and other research of the Slovak language. Translations also occur in the category of journalistic texts, but their identification is substantially more problematic, sometimes even impossible. For instance, translations of agency news provide no indication that the text has been translated. Such information cannot be collected automatically.

1.1. Corpus structure

The Slovak National Corpus provides several specialised subcorpora. The main corpus, called *prim* (primary, general corpus), is offered with several options, depending on the target audience. The option can be seen as constructed according to a matrix where each column selects the audience allowed to access the corpus (according to the license obtained), and each row selects the type of texts in the corpus.

In Table 1, *public-* denotes corpora publicly accessible, *juls-* corpora that are accessible only from within the premises of the Ľ. Štúr Institute of Linguistics (for anyone interested in the research), *snk-* corpora accessible only to employees and collaborators of the Slovak National Corpus department; *-all* is the corpus containing all the texts collected, *-sane* has some restrictions on the texts included (no texts without bibliographic annotation, no texts published by ethnic Slovaks abroad, no texts without proper diacritics), *-img* contains only fiction (both original Slovak and translations), *-inf* only informational texts, such as newspaper articles, *-prf* professional texts (scientific and popular science texts), *-sk* texts that are not translations from another language, *-skimg* original Slovak fiction. Balanced corpus *-vyv* contains ⅓ fiction, ⅓ professional texts and ⅓ informational texts; *55-89* texts that originated between the years 1955 and 1989; *mak* is the manually morphologically annotated corpus.

prim	**public-**	**juls-**	**snk-**
-all	719	729	770
-sane	699	709	734
-img	99		
-inf	515		
-prf	82		
-sk	509		
-skimg	317		
-vyv	247		
55–89	44		
mak	1.2		

Table 1. *Size of subcorpora (in millions of tokens)*

1.2. Annotation

The entire SNK is lemmatised and morphologically annotated with detailed bibliographic and style-genre annotation. A small part is manually annotated syntactically. With regards to the processing of text annotation in the Slovak National Corpus, we were inspired by the EAGLES recommendations and tried to follow the principles defined by G. Leech (1993):

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
2. It should be possible to extract the annotations by themselves from the text.
3. The annotation scheme should be based on guidelines which are available to the end user.
4. It should be made clear how and by whom the annotation was carried out.
5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
7. No annotation scheme has the a priori right to be considered as a standard.

From a linguistic point of view, for each type of annotation we are searching in particular for a compromise with point 6 of Leech's maxims of annotation. There was a great deal of discussion about the quality and quantity of the meta-information in the corpus and we carried out several revisions. The bibliographical annotation records not only included the author, work title, place and year of issue, but also the sex of the author, language of the text, etc. In the case of translations, there was also included the name and gender of the translator, the original version of the author's name and the original language of work. Style-genre annotation is extended to domains (e.g. humanities, sciences, law, politics, technology, economics) and subdomains (e.g. for area belief and the supernatural there are 3 subdomains: 1. religion, faith, sect; 2. theology; 3. supernatural, occult, magic, astrology). A special item represents the medium from which the text was obtained: book, newspaper, magazine, textbook, internet, manuscript, etc.[1] It is important to have various text styles and genres in the corpus with their precise metadata annotation, especially either for the preparation of the new dictionary of the Slovak language, or for the creation of a frequency dictionary of Slovak and for the Slovak Terminology Database. The collection of texts (mainly in the areas of technical and natural sciences) is obviously determined by the following factors:

– new scientific production in specific domains is more frequently written in foreign languages than in Slovak;
– older scientific works are often considered obsolete and not relevant even from the point of view of terminology, and their authors cannot make them available for any purposes.

The greatest differences from classic, non-corpus linguistics are found in morphological, respectively morphosyntactic annotation. The Slovak tagset[2] is based on formal morphology and the attributive basis (similar to MultextEast). 10 kinds of traditional parts of speech were extended to 19 word classes, covering all tokens that can occur in the texts: **S** – nouns, **A** – adjectives, **P** – pronouns, **N** – numerals, **V** – verbs, **D** – adverbs, **E** – prepositions, **O** – conjunctions, **T** – particles, **J** – interjections, **G** – formal participles, **R** – reflexive morphemes *sa/si*, **Y** – conditional morpheme *by*, **0** – numbers, **W** – abbreviations and symbols, **Q** – unclassifiable parts of speech, **%** – citation, **Z** – punctuation, **#** – non-word elements. Several linguistic units and events had to be treated in a special way, such as secondary prepositions which are annotated as a preposition – noun – preposition (*v*[E] *súvislosti*[S] *s*[E] – 'in connection with'). Inflectional paradigms of some parts of speech are reflected in the second position of the attribute

[1] For more details about the style/genre annotation see http://korpus.juls.savba.sk/bibstyle.html

[2] http://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf

(e.g., pronouns may have the inflectional paradigm of nouns, adjectives, adverbs, pronouns, mixed and incomplete).

The manually morphologically annotated corpus has 1.2 million units and is used for training tools for automatic annotation and as a reference source of statistical information about the Slovak language. The following graph shows the total share of prepositions, conjunctions, adverbs, particles and interjections in the manually annotated corpus and the number of their unique occurrences. A preposition is found in these texts, on average, 795 times, conjunction 690 times, adverb 14 times, particle 127 times and interjection 4 times. Interjections in written texts are generally very rare (in fiction, more often in translations than in the original Slovak works). Adverbs have the most comprehensive inventory.

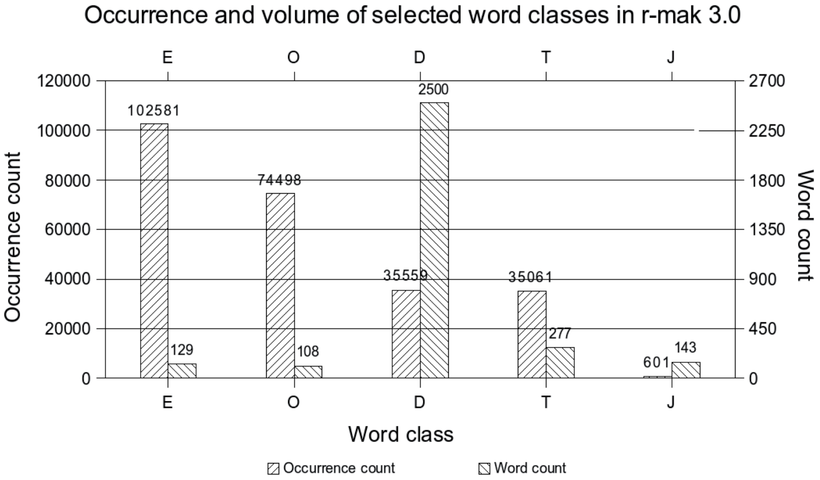


Figure 1. *Number of total occurrence count (left axis) and unique words (right axis) of selected word classes in r-mak 3.0*

Manual syntactic annotation (its analytical layer) of SNK texts is implemented with the use of technical tools along with the linguistics and technical manual of the Faculty of Mathematics and Physics of Charles University in Prague. The first phase included a double syntactic annotation of the texts previously manually morphologically annotated. The Slovak Dependency Treebank has a wide coverage of styles and genres, including novels, fairy tales, journalistic texts, scientific articles and historical texts. These texts were independently annotated by two different annotators. There are about 50,000 sentences in the treebank, each annotated twice by two different annotators. Further plans include comparing the annotations, pinpointing the differences (presumably due to errors and different views on syntactical structure) and creating a corrected version of the treebank.

1.3. Corpus access

The corpus has been indexed with the use of Manatee. There are two possibilities to make query in the corpus. The first is based on using the multiplatform (Tcl/Tk) *Bonito* client to access the *Manatee* server using its own TCP/IP protocol. This approach provides users with complete access to all the advanced querying, sorting and statistical features of the server although requires the installation of specialised software. The other possibility is to use simple web-based access where only basic features are present. In both cases, the search interface provides CQL compatible query syntax. However, in the last few years, the ability of an average user to install arbitrary software (and use anything that is not web-based) has declined considerably, and new corpus users often face an insurmountable obstacle of downloading, unpacking and running the *Bonito* client. Due to this fact, we are planning to transfer the corpus to *Manatee-2*, which provides a web-based interface as a suitable replacement for the Tcl/Tk client.

2. Corpus of Spoken Slovak

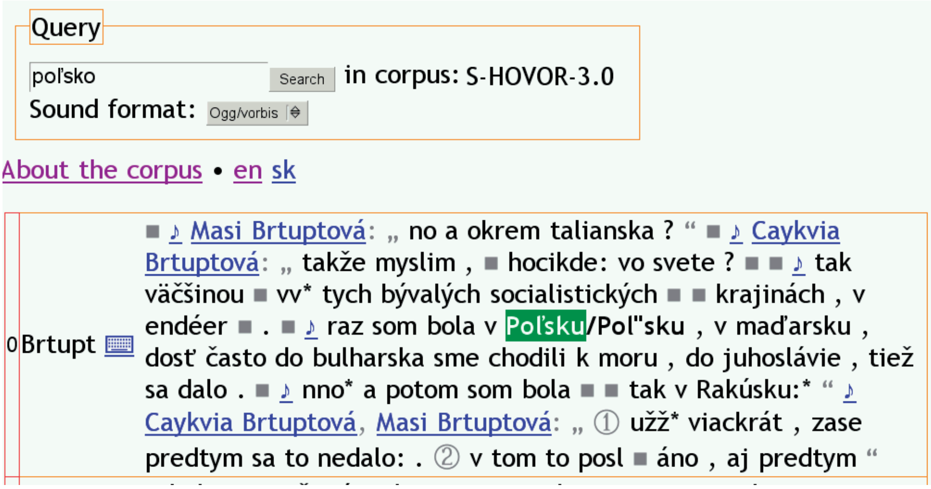
The corpus of Spoken Slovak is a project focused on recording a reasonable amount of sound samples from contemporary Slovak together with manual phonemic transcription, automatic lemmatisation and morphosyntactic analysis. At the time of writing, the corpus contains about 1.65 million tokens (punctuation included). This includes spontaneous dialogue, guided conversations, samples of talk shows from radio and television, lectures, and sermons from all over Slovakia. A testimony of people who were persecuted during previous regimes (from the Oral History Project of the Nation's Memory Institute) represents a significant part of the recordings.

The archive is kept in FLAC format and the entire recordings have been converted into Ogg/Vorbis and Ogg/Speex formats (for easier handling and transcription). In order to finally link the files throughout the corpus web interface we divided them into small chunks corresponding to the relevant dialogue turns. The source archive currently amounted to over 200 GB.

One of our primary goals was to make this corpus unencumbered by the usual copyright and privacy concerns that plague similar projects. We must not only be aware of copyright law, but also the law on the protection of personal data. We undertook this by removing any sensitive information (personal names) before including the recordings in the archive, and by including only those recordings where we had explicit expression of consent by all the relevant participants to include the recordings in our archive.

For transcription, we used *transcriber* (Barras et al. 2001), with a detailed set of tags to annotate both internal speech features and external sound events influencing the recorded discourse.

The corpus can be accessed in two independent ways. One of them uses the standard *Bonito* client in the same way as the preferred access to the main Slovak National Corpus. The other uses a specialised web interface which offers additional visual representation of transcription and annotation as well as links to the sound recordings themselves.



Query

poľsko Search in corpus: S-HOVOR-3.0

Sound format: Ogg/vorbis

About the corpus • en sk

0Brtupt ⌨ ■ ♪ Masi Brtuptová: „ no a okrem talianska ? “ ■ ♪ Caykvia Brtuptová: „ takže myslim , ■ hocikde: vo svete ? ■ ■ ♪ tak väčšinou ■ vv* tych bývalých socialistických ■ ■ krajinách , v endéer ■ . ■ ♪ raz som bola v Poľsku/Pol"sku , v maďarsku , dosť často do bulharska sme chodili k moru , do juhoslávie , tiež sa dalo . ■ ♪ nno* a potom som bola ■ ■ tak v Rakúsku:* “ ♪ Caykvia Brtuptová, Masi Brtuptová: „ ① užž* viackrát , zase predtym sa to nedalo: . ② v tom to posl ■ áno , aj predtym “

Figure 2. *Spoken corpus web interface*

3. Collocation dictionary

The collocation dictionary is the first of its kind in Slovakia and is aimed at the registration and description of multi-word lexemes and phrasemes as well as typical collocations with wide collocability which are frequentially differentiated and so limited in that way (Majchráková, Ďurčo 2010). The dictionary provides an overview of the combinatorial behaviour of words, primarily the most frequent nouns from the Slovak National Corpus. The combinatorial potentials of words are the basis for the creation of so-called collocational templates which are the basis for the patterns of collocations (Ďurčo 2007). We use the Mediawiki system to access and edit the database, utilising its advantages such as simple editing syntax, multi-editing, history of editing, instant updating, forum discussion (which provides interaction between editors), verification system of version changes and others. In the initial phase of the project we used the data from the Slovak National Corpus in the version prim-3.0 (containing about 350 million tokens). As the SNK database was expanding, for further work we decided to use the version prim-4.0 containing 550 million tokens. The question was whether added data would change statistical measures in so far that it could reflect a different collocational profile of a word and if it would markedly change the word's combinatorial preferences. This issue was solved by creating contrastive profiles of a given word using a different amount of data. Based on the new version *prim-5.0* we are currently preparing a dictionary of adjective collocations.

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 context, obtain statistical information like absolute frequency, MI-score, t-score, MI-score, MI3, log likelihood, min. sensitivity and salience to recognize word co-occurrences. Another highly specialised tool for an efficient analysis of relevant collocations is Sketch Engine,[3] which is used internally at our institute.

While the basic criteria for searching collocations is frequency, not all statistically significant collocations are automatically included in the collocational profile of the entry (for example, proper names and terms have to be excluded). On the other hand, statistically less significant or less frequent combinations of words can also be marked as fixed or relevant. One of the reasons is that the main SNK database (*prim-4.0*) used for this purpose is not a well-balanced corpus which predominantly consists of of journalistic texts. The fact that many collocations come from this field may imply that word partnership found in journalistic texts implies that many typical combinations of words from this field are usually frequent.

4. Slovak WordNet

There is currently an ongoing effort in collaboration with the Technical University of Košice in building a basic Slovak WordNet database. We plan to use the database as a skeleton of a basic English-Slovak-German-Polish-Lithuanian

[3] http://www.sketchengine.co.uk/

dictionary[4] project. The building process consists of mapping automatically generated Slovak synsets to English synsets from WordNet v.3.0. The synset generation has been described in (Genči 2009); the synsets are manually corrected before being added to the database. We use special annotation to mark synsets that do not have a clear English equivalent. Our goal is to build synsets containing the ten thousand most frequent words from the Slovak National Corpus (nouns, adjectives, verbs and adverbs), together with a complete set of their hypernyms (i.e. each Slovak synset will have a hypernym, unless mapped to those few English synsets that do not have a hypernym).

5. Slovak Terminology Database

The Slovak Terminology Database is a monolingual database provided with both conceptual and linguistic information. We drew inspiration especially from the ISO 10241:1992 International terminology standards – Preparation and layout. The EUROVOC 4.2 Thesaurus was chosen as the classification system.

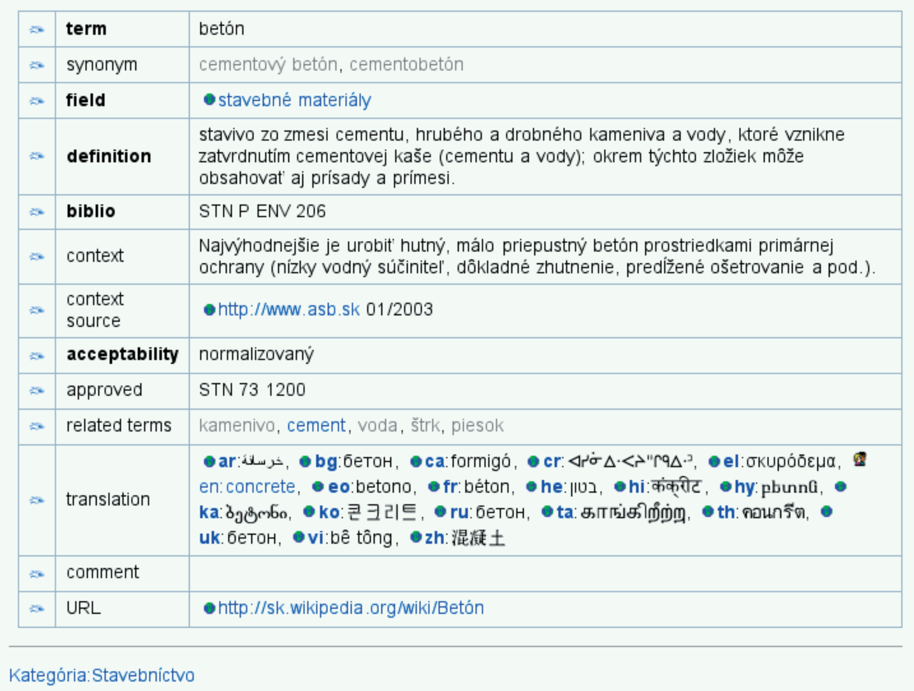
The project methodology has adopted a textual approach to the terminology extraction of lexical units, potential terminological units from running specialised texts and identification of the concept they refer to.

The term record comprises 11 data categories, 7 out of which are obligatory. In order to satisfy the needs of professionals, the lay public and – last but not least – the translation and interpreting public, obligatory categories include: definition, domain, context, related terms and sources of both definition and context. The remaining 4 optional fields of the term record feature: synonym, foreign language equivalent, comment and links to relatively reliable web pages likely to remain online in the foreseeable future (such as other terminology databases, government resources or Wikipedia pages).

The terminology database covers the following domains: Astronomy, Security and Law, Migration Policy, Construction, Corpus Linguistics, Phraseology, Phonetics and Phonology, Bilingualism, Civil Security, Historical Linguistics, Historical Units of Measure, History, Social Fire Protection, Social Work, Volcanology, Game of Chess.

The database is implemented as a MoinMoin wiki system with a customised parser displaying information in a tabular format where each table row corresponds to one data category and MoinMoin categories corresponding to terminology domains.

[4] As part of the Slovak Online project – Lifelong Learning Programme 504873-2009-LLP-SK--KA2-KA2MP



term	betón
synonym	cementový betón, cementobetón
field	● stavebné materiály
definition	stavivo zo zmesi cementu, hrubého a drobného kameniva a vody, ktoré vznikne zatvrdnutím cementovej kaše (cementu a vody); okrem týchto zložiek môže obsahovať aj prísady a prímesi.
biblio	STN P ENV 206
context	Najvýhodnejšie je urobiť hutný, málo priepustný betón prostriedkami primárnej ochrany (nízky vodný súčiniteľ, dôkladné zhutnenie, predĺžené ošetrovanie a pod.).
context source	● http://www.asb.sk 01/2003
acceptability	normalizovaný
approved	STN 73 1200
related terms	kamenivo, cement, voda, štrk, piesok
translation	● **ar**:خرسانة, ● **bg**:бетон, ● **ca**:formigó, ● **cr**:ᐊᓯᓃᐃᐧᐸᐢ"ᒋᑫᐃᐧᐣ, ● **el**:σκυρόδεμα, en:concrete, ● **eo**:betono, ● **fr**:béton, ● **he**:בטון, ● **hi**:कंक्रीट, ● **hy**:բետոն, ● **ka**:ბეტონი, ● **ko**:콘크리트, ● **ru**:бетон, ● **ta**:காங்கிறீற்று, ● **th**:คอนกรีต, ● **uk**:бетон, ● **vi**:bê tông, ● **zh**:混凝土
comment	
URL	● http://sk.wikipedia.org/wiki/Betón

Kategória:Stavebníctvo

Figure 3. *Example of a terminology entry*

6. Conclusion

In its current form, the Slovak National Corpus provides the basic research material for all categories of users and anybody who is interested in the Slovak language. It is a basis for the creation of orthographic or grammar manuals, a basis that is readily accessible via the Internet and essentially provides wider potential within the framework of the automated processing of large numbers of real life texts. Recently, a significant proportion of users came from a group of teachers in primary and secondary schools who in their career development discovered the possibility of using SNK databases in the teaching process. After completing the second phase of the SNK project in 2011, information on the corpus data and representation of the contemporary Slovak language should be available in the form of a frequency and collocation dictionary. We hope that the next phase will also include a continuation of the work on building and balancing the primary national corpus and linguistic annotation, the spoken corpus and parallel corpora, the Slovak Terminological Database as well as Slovak WordNet. New special projects will be implemented in accordance with availability of funding and the requirements of users.

References

Barras, C., Geoffrois E., Wu, Zh., Liberman, M., 2001: *Transcriber: Development and use of a tool for assisting speech corpora product*, "Speech Communication", 33/2001, pp. 5–22.

Ďurčo, P., 2007: *Collocations in Slovak (Based on the Slovak National Corpus)*, [in:] *Computer Treatment of Slavic and East European Languages*, eds. J. Levická, R. Garabík, Bratislava, pp. 43–50.

Genči, J., 2009: *Synset Building Based on Online Resources*, [in:] *NLP, Corpus Linguistics, Corpus-Based Grammar Research*, eds. J. Levická, R. Garabík, Bratislava, pp. 106–113.

Leech, G., 1993: *Corpus Annotation Schemes*, "Literary and Linguistic Computing", Vol. 8, No. 4, pp. 275–281.

Majchráková, D., Ďurčo, P., 2010: *Compiling the First Electronic Dictionary of Slovak Collocations*, [in:] *LEXICOGRAPHICA. Feste Wortverbindungen und Lexikographie. Kolloquium zur Lexikographie und Wörterbuchforschung*, ed. P. Ďurčo, Berlin, pp. 105–114.

Páleš, E., 1994: *SAPFO. Parafrázovač slovenčiny*, Bratislava.

Slovak National Corpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2011. Available from WWW: http://korpus.juls.savba.sk/

Zákon č. 428/2002 Z. z. *o ochrane osobných údajov. Národná rada Slovenskej republiky 2002, 2004, 2005*, [in:] *Zbierka zákonov Slovenskej republiky*, Bratislava.

Słowacki Korpus Narodowy i jego zasoby językowe

Streszczenie

Artykuł opisueje projekty związane ze Słowackim Korpusem Narodowym. Pierwszy z nich to wielki reprezentatywny korpus współczesnego pisanego języka słowackiego (od reformy ortografii z roku 1953). W chwili obecnej korpus zawiera ponad 770 milionów słów i liczba ta wciąż rośnie. SNK składa się z kilku podkorpusów reprezentujących beletrystykę, teksty profesjonalne, teksty prasowe oraz zbiór zrównoważony.

Teksty są automatycznie anotowane morfologicznie oraz za pomocą innych metadanych. Korpus jest dostępny dla wszystkich na zasadach niekomercyjnych. Interfejs wyszukiwarki umożliwia wyszukiwanie morfologiczne i składniowe, bogate analizy statystyczne oraz wyszukiwanie kolokacji. Inne korpusy związane z SKN to: anotowany ręcznie korpus treningowy wielkości 1,2 mln słów; korpus języka mówionego, zawierający obecnie ponad 1,5 miliona słów oraz bank drzew składniowych zawierający 50000 zdań. Innym projektem w ramach SNK jest terminologiczna baza danych.