The **Slovak National Corpus Department** (SNC) is the youngest and most modernly equipped department of the Linguistics Institute, and the only one of its kind in Slovakia. The department was founded in 2002 with the support of the Ministry of Culture of the Slovak Republic (program for maintaining the national language), the Ministry of Education (informatization and use of innovative methods in teaching) and the Slovak Academy of Sciences (SAS). An eight-member team of predominantly young scientists is involved in the project: **Construction of the Slovak National Corpus and the electronization of linguistic research in Slovakia.** Thus, a new field of research, **corpus linguistics**, was established in Slovakia and is developing successfully in the department. In 2005 the members of department were awarded the **SAS Prize for construction of scientific infrastructure**.
organized the biennial international conference *Slovo* on natural language processing and corpus linguistic research, with participation by Slovak as well as foreign researchers (from Austria, Bulgaria, Croatia, the Czech Republic, France, Germany, Hungary, Poland, Russia, Slovenia, Spain, Ukraine, etc.). The published conference proceedings contain contributions on the preparation, investigation, and results of diverse national and international projects in the field of construction and use of general and specific corpora and databases, in the field of language analysis and synthesis, automatic translations, computer lexicography and terminography, e-learning etc.

### What is a linguistic corpus and what is it not?

A corpus is a collection of texts – words and sentences in real use – available in electronic form in a unified format accompanied by language information and equipped with powerful search tools.

A corpus is not an electronic library. Individual texts (books, articles) cannot be downloaded or read as a whole. One can read only short contexts of a desired word, word phrase or grammatical units chosen out of all corpus sources. It is not a text archive, although within the framework of the corpus the texts are stored in their original format, but only internally, serving for the initial processing and as a structured collection of original documents. It is not even an explanatory dictionary or a codifying orthographic reference book with authoritative orthographic rules, although the corpus is a source of material for the preparation of the above-mentioned or other linguistic works.

### What texts are contained in the Slovak National Corpus?

The primary, general corpus *prim* covers Slovak texts which arose after the year 1955. The texts are preserved in the original version, no language corrections are made. Three major styles are represented in the corpus: journalistic, literary, and professional (popular non-fiction) as well as various other genres and areas. The corpus database comprises texts from throughout Slovakia as well as texts by Slovaks living abroad, texts originally in Slovak and translated from other languages. For specialized research, the general corpus *prim-*all can be divided into independent subcorpora:

- **sane** – does not contain linguistic texts, texts without diacritics, texts from Slovaks living abroad etc.
- **yyv** – journalistic, literary, and professional texts are represented by a third share each
- **inf** – journalistic texts only
- **prf** – professional texts only
- **img** – literary texts only
- **skimg** – originally Slovak literary texts only

The corpus version available on the Internet contains only texts obtained from providers under a license agreement.

### Who are the text providers?

Authors, translators, publishers – the owners of copyright or distribution rights who are asked by the members of the SNC department to provide materials. A portion of the texts are obtained under the law providing for compulsory copies of periodicals, non-periodic publications, and copies of audiovisual works ([http://juls.savba.sk/vytlacok.html](http://juls.savba.sk/vytlacok.html)). The use of the texts in the Slovak National Corpus is governed by the provisions of the Copyright Act. We thank all 600 [providers](http://korpus.sk/contributors.html) for their cooperation.

### How are the corpus texts processed?

For texts acquired in electronic form, the marks and symbols of the editors and programs (in which they were created) as well as the graphical elements (charts, graphs, pictures) are removed and the plain texts are converted to a unified format. Texts existing in non-electronic form are scanned, recognized and transferred to a plain text. Then the texts are corrected to be made consistent with the printed version. Afterwards they are also converted to a unified format. Each text is provided with external annotation – bibliographical information, style and genre characteristics. In the next phase, all the texts are segmented – divided into separate units (words, numbers, punctuation marks, signs, symbols) and linguistically tagged.
What language information is included and provided in the Slovak National Corpus?

The corpus texts and text units are accompanied by:

- external, bibliographical, style, and genre annotation (http://korpus.sk/bibstyle.html) – each text is accompanied by information about its origin (author, translator, title, place and year of publication, publisher), about the style and genre classification (novel, poem, monograph) and about the field (law, technology, economics),
- internal, morphological or morphosyntactic annotation (http://korpus.sk/morpho.html) and lemmatization – each word is given information about its grammatical categories (parts of speech, case, number, tense, person) as well as about its base (dictionary) form.

Thanks to this annotation desired words and phrases can be looked up in the corpus and the subcorpora:

- according to a specific word form: slovenskými – the form slovenskými is displayed in all contexts,
- according to the base form – lemma: slovenský – contexts of the word slovenský are displayed in all its forms,
- according to a morphological characteristic – tag: AAis7 – contexts of all the adjectives (A) are provided with the adjectival declension (A), masculine inanimate (i), singular (s), 7th case (instrumental case),
- with the help of regular expressions (a collection of signs and special symbols; http://korpus.sk/usage.html) one may search for words or parts of words containing specific combinations of signs: .*skými (dot, asterisk, skými) – all the word contexts ending in -skými are displayed (ľudskými, ženskými, školskými).

These types of searching can also be used when working with the corpora of other languages (http://korpus.sk/links.html).

Who and what is the Slovak National Corpus intended for?

- General public, authors, translators, professionals working with language to verify the word meaning or orthography, word collocability, occurrence in phrases, terminological combinations, synonym sets, and so on.
- Teachers and foreign lecturers for independent, creative, and dynamic work when teaching Slovak as a mother tongue or teaching Slovak for foreigners to prepare their own language and stylistic activities.
- Lexicographers information base for creating new monolingual and bilingual dictionaries.
- In various grammatical, stylistic, contrastive and other linguistic researches to research into the real functioning of language and the dynamics of language.
- In mathematical and statistic analyses, in philosophy, cultural studies, sociology, speech therapy, neurology, psychology, and didactics language information from the corpora.
- In the field of natural language processing (NLP) to create different computer tools, automatic translations, and so on.

Corpus of Spoken Slovak

The database of the Corpus of Spoken Slovak (http://data.juls.savba.sk/oral/) contains audio records of spontaneous and semi-prepared speech from the entire Slovak territory and their text transcripts. Specific characteristics of spoken language are selectively captured in the transcripts, such as irregular structure of an utterance, pronunciation variants, means of speech modulation, and the presence of non-linguistic elements. The Corpus of Spoken Slovak provides material for research and description of the real form of contemporary standard spoken Slovak.

Slovak Terminology Database

The Slovak Terminology Database (http://data.juls.savba.sk/std/) includes Slovak terms and relevant terminological information from various disciplines. The database serves for the stabilizing and unification of terminological systems, through the cooperation of linguists and experts from relevant professional fields. Its aim is...
to summarize and standardize the Slovak terminology. The terms are obtained from the specialized professional corpora (law texts, economy texts, etc.).

Parallel corpora

In parallel corpora a single text is displayed in two or more languages – mutual translations or translations of the same text from a third language. Research into language features and devices in parallel corpora enables:

- knowledge of the real functioning of the translation equivalents
- knowledge of the mutual relationships of specific languages
- refinement of contrastive research
- improvement and development of the entries in bilingual dictionaries
- utilization in translation practice and in education of translators
- utilization in teaching foreign languages or Slovak as a foreign language


Collocational Dictionary of Slovak. Substantives

300 collocative profiles of the most frequent Slovak substantives have been processed under the auspices of the project VEGA The Confrontational Research of Collocations in Slovak and German. These are phrases with a substantive, verb, adjective, and so on occurring in various forms. The summary of multi-word lexicalized and phraseological expressions, typical and common textual expressions is a good source of semantic relationships in a vocabulary as well as an associative tool for different usage in practice or theory.

Frequency parameters of contemporary Slovak

Contemporary corpus databases contain from hundreds to thousands of million of units (words, marks, numbers, punctuation marks). Through analysis of these huge information sources the systematic and structural features of a specific language and language means functioning are detected, frequency dictionaries or lists are produced (http://korpus.sk/stats.html) and typologically related or unrelated languages are compared.

A typical feature of Slovak is the frequent usage of vowels. The most common letters are: o, a, e, n, i, r, t, s, v, l, among the most predominant letter pairs the following consonant pairs most frequently occur: ov, st, pr, ne, po, to, na, ch, en, ie.
**Linguistic sources**

The most renowned and usable product of linguistic research electronization in Slovakia is a free file of the electronic forms of up-to-date editions of the Short Dictionary of Slovak, Rules of Slovak Orthography, Slovak Synonym Dictionary, Dictionary of Foreign Words and older Dictionary of Slovak (http://slovniky.korpus.sk) as well as other dictionary works, monographs, miscellanies and linguistic periodicals (http://korpus.sk/dicts.html). On average, **40 000 input queries** are searched a daily in the database.

**International projects and cooperation**

**Mondilex**

From 2008 to 2010 the researchers of SNC department were involved in the Seventh Framework Programme focused on building a research infrastructure for the digital resources in the Slavic lexicography. Partners from Bulgaria, Poland, Russia, Slovenia, and Ukraine also participated in the project.

**Slovak Online**

Slovak Online (http://slovake.eu/sk/) is a project supported by the European Commission within the framework of the program Grundtvig – Lifelong Learning Programme. It is a free Slovak language course suitable for foreigners, tourists, and inhabitants of the border areas, or students interested in learning Slovak, as well as for Slovaks living abroad to keep in touch with their mother tongue, Slovakia, and its culture. The main contribution of the project participants from the Slovak National Corpus was the creation of a multilingual dictionary.

**EuroMatrix+** is a project of the Seventh Framework Programme – Information and Communication Technologies. Implementing the Slovak language with a machine translation system is the focus of attention of the SNC. As part of the project, sample systems for each official language in Europe are being created and machine translation developers are provided with the infrastructure for building statistic models of translation. The project is aimed at creating scientifically new possibilities and methods related to current technological needs.

**CESAR** is a project supported by the European commission through the ICT Policy Support Programme (ICT PSP). Its target is to make existing linguistic sources available to a wide range of volunteers and to contribute to the creation of cross-European digital market. The target community of language sources and services users includes stakeholders involved in the modern digital market: common users, professional users (in the field of trade, administration, media, education, libraries) and experts (researchers, industrial workers, politicians etc.). Partners from Hungary, Poland, Croatia, Serbia, Bulgaria and Slovakia are involved in CESAR as well as other European countries in META-NET. The task of the Slovak National Corpus is to make available a comprehensive set of Slovak language sources.

---

The most frequent substantives, adjectives and verbs in the written and spoken corpora and their positioning among the top 100 most frequent units

<table>
<thead>
<tr>
<th>prim-5.0-all</th>
<th>s-hovor-3.0</th>
<th>prim-5.0-all</th>
<th>s-hovor-3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovensko 79.</td>
<td>vec 88.</td>
<td>Slovensko 79.</td>
<td>vec 88.</td>
</tr>
<tr>
<td>deň 82.</td>
<td>dieta 89.</td>
<td>deň 82.</td>
<td>dieta 89.</td>
</tr>
<tr>
<td>čas 91.</td>
<td>byť 1.</td>
<td>čas 91.</td>
<td>byť 1.</td>
</tr>
<tr>
<td>strana 96.</td>
<td>mat 21.</td>
<td>strana 96.</td>
<td>mat 21.</td>
</tr>
<tr>
<td>spoločnost 100.</td>
<td>vediet 34.</td>
<td>spoločnost 100.</td>
<td>vediet 34.</td>
</tr>
<tr>
<td>byť 1.</td>
<td>ist 37.</td>
<td>byť 1.</td>
<td>ist 37.</td>
</tr>
<tr>
<td>povedať 66.</td>
<td>môcť 40.</td>
<td>povedať 66.</td>
<td>môcť 40.</td>
</tr>
<tr>
<td>môcť 40.</td>
<td>hovoriť 44.</td>
<td>môcť 40.</td>
<td>hovoriť 44.</td>
</tr>
<tr>
<td>musieť 76.</td>
<td>dat 57.</td>
<td>musieť 76.</td>
<td>dat 57.</td>
</tr>
</tbody>
</table>

The entry book (kniha) in the Slovak-English-German-Polish-Lithuanian dictionary.