Slovak National Corpus Project (Slovenský národný korpus)

The Slovak National Corpus (SNC) is a research project focused on constructing a computer based corpus of texts of the contemporary Slovak language (1955-2005) with an orientation towards written texts. Later it will be enlarged with texts from other periods and spheres of Slovak language usage. The Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences, which began construction of the SNC in 2002, is thus putting into motion an important phase of computerization of linguistic research in Slovakia and is preparing conditions for computer processing of the Slovak language as a natural language.

Worldwide, corpus linguistics already has a more than forty year tradition. In Slovakia, this discipline began to develop ten years ago with very limited technical and personal facilities. In 2002, the Department of the Slovak National Corpus arose as a specialized department of the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences in Bratislava with the aim of building and administering an automatized text database of the Slovak language in the widest possible range.

The Slovak National Corpus is built as a general monolingual corpus, which in the first phase (year 2003) is compiling written texts originating in the years 1990-2003 in a range of about 30 million words with lemmatization, morphological and source (bibliographical and style-genre) annotation. In the second phase (up to 2006) the representative span of the written texts will be enlarged to other periods of the contemporary language (1955-2005) to the range of 200 million words and its selected sample will be syntactically annotated. At the same time specific sub-corpora of diachronic and dialectological texts will be built, as well as a terminological and lexicographical database.

The Slovak National Corpus is provided primarily for lexicographers (dictionary production), for grammatical and stylistic research (grammatical and orthographical handbooks; varieties of the national language and their communicational application). We assume that it will also find its use in schools (preparing of orthographical, grammatical and stylistic textbooks; teaching Slovak as a foreign language). Prospective specific sub-corpora of historical and dialectological texts will be viewed as a preservation and a long-term, widely accessible source for an important part of our cultural heritage.

You can search the test version of the Slovak National Corpus ("prim0.1") using the corpus manager "Manatee", via the client "Bonito". The use of the corpus manager for access to the texts included in "prim0.1" is achieved through your own account which will be assigned to you after sending us a completed and signed form available through our website. This form should be sent to the contact address of the SNC. Both the corpus manager and the client can also be downloaded from our website. The publicly accessible corpus "prim0.1-public" can be used for simple queries via the www interface. The composition of texts is identical with the original corpus "prim0.1". This public version is

provided for anyone interested in knowing the functioning of contemporary Slovak language, but it doesn't supply codification and grammar handbooks or bibliographical annotation.

The Slovak National Corpus website can be found at <http://korpus.juls.savba.sk>. The contact coordinates for the project are: Slovenksý nárdoný korpus; Jazykovedný ústav Ľudovíta Štúra SAV; Panská 26; 813 64 Bratislava, Slovakia; Tel: +421 (2) 5441 0304; Fax: +421 (2) 5441 0307; E-mail: korpus@juls.savba.sk

Mária Šimková, Director of the Department of the Slovak National Corpus