Slovak language in computer processing

Radovan Garabík
Mária Šimková
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Web-based interface to Manatte/Bonito platform running Slovak National Corpus

General information

Slovak language belongs to the West Slavic language group, together with Polish, Czech and Sorbian languages. It remained especially close to the Czech language, due to close historical and linguistic ties between the languages (especially in the former common country). Slovak is a typical Slavic language in retaining complex inflectional and derivational morphology, with only minor simplifications compared with neighbouring languages.

Slovak language is spoken by about 5 million people in Slovakia, but also in some other countries. The most numerous Slovak community is in the USA (1 million people, many of them actually speaking the language), Czech Republic, smaller communities are present in Romania, Hungary, Serbia and other countries.

Slovak orthography is mostly phonemic, with noticeable etymological and morphological features. The language is written in the Latin script, with acute accents marking (phonemic) vowel length, há?eks marking palatals and post-alveolar fricatives. The latest substantial orthography reform has been implemented in 1953, when the language gained contemporary form practically in all of its aspects.

Major linguistic research centres in Slovakia

The *Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava* is the central linguistic institution in Slovakia. Its main area of research is traditional linguistics, with the emphasis on (but not limited to) Slovak language, its history, dialectology, etymology and lexicography, as reflected in a sizeable amount of dictionaries produced. Traditionally, the Institute was connected with the task of regulating the Slovak language, defining its grammar and orthography and producing prescriptive dictionaries. However, in the last decades the main orientation shifted more towards general linguistic research (not limited to the Slovak) and descriptive or bilingual dictionaries and other popular and scientific publications in the field of sociolinguistics, onomastics, language theory, lexicography and others.

Institute of Informatics, Slovak Academy of Sciences, Bratislava is dealing with theoretical and applied research in the field of computer science, information technologies and artificial intelligence. Their NLP tools include information retrieval, knowledge representation speech synthesis and recognition. The main research direction of *Faculty of Arts, University of Prešov* in Prešov is phonetics, phonology, derivatology and morphematic structure of the Slovak. The results of their research are often used in Slovak NLP. An extensive research of child speech is carried out in cooperation with the Faculty of Education, especially in international projects. *University of SS Cyril and Methodius, Trnava* puts special emphasis on German-Slovak confrontational research on collocations, valency and corpus lexicography as well as phraseology and paremiology. *Faculty of Humanities, Matej Bel University Banská Bystrica* is active in sociolinguistics and research in communication and stylistics. Part of their database of recordings of spontaneous speech in the city of Banská Bystrica and surroudings become part of the Corpus of Spoken Slovak. *The Technical University of Košice* is active especially in NLP, computer aided lexicography and ontology research.

Slovak language in computer processing

The *Slovak National Corpus department* was established as a special project of the Ministry of Culture and Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences in 2002. This marked rapid expansion of Human Language Technology research of the Slovak language, with the Ľ. Štúr Institute of Linguistics becoming the leading research institution in the field. The Institute actively leads R&D in computational linguistics oriented towards contemporary written and spoken Slovak language, covering all aspects of language analysis and processing. Its strong IT background is demonstrated in the number of tools and resources that the department developed during its existence, among the most important are the Slovak National Corpus database and all the necessary associated tools.

The Slovak National Corpus is a big, representative corpus of modern written Slovak (since the 1953 orthography reform). Currently, the whole corpus contains about 780 million tokens. There are several specialised subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus, texts written before 1989). The corpus is automatically lemmatised and morphologically annotated, using its own tagset and morpfology analyser. Access to the corpus is provided free of charge, for non commercial purposes. The corpus legal status is rather unusual, if compared with other language corpora – The Ľ. Štúr Institute of Linguistics did obtain license agreements to use the texts in building the corpus database (for non commercial, research and educational purposes), and the corpus therefore includes the texts with full legal compliance. *New Dictionary of Contemporary Slovak* (*Slovník súčasného slovenského jazyka*) is being compiled with the help of the corpus and is the first Slovak dictionary based predominantly on corpus resources.

The department actively works on several smaller, but no less important projects. *Corpus of Spoken Slovak* is a representative corpus of standard spoken Slovak as spoken throughout Slovakia, consisting of about 160 hours of recordings (1.6 million words). The recordings are manually transcribed on orthographic and phonemic level. Parallel corpora include Slovak-Czech, Slovak-Russian, Slovak-French, Slovak-Bulgarian and Slovak-English corpus. The texts (mostly fiction translations) are automatically sentence-aligned and morphosyntactically tagged. Slovak dependency treebank contains about 50,000 manually syntactically annotated sentences.

The *Slovak Terminology Database* focuses on the field of law, economy and technology, offering about 5000 terminological records that can be classified by circa 20 EUROVOC descriptors corresponding to different soft and hard sciences.

International collaboration

With expansion of data and resources, it is more and more apparent that standardisation and availability of resources, documentation and interoperability are important for further scientific research and technology deployment. Since the beginning, the Slovak National Corpus department is committed to making available all the resources and tools under favourable Open Content and Open Source licensing policies (if other copyright restrictions permit). Collaboration with other partners has proved to be very fruitful and we expect to extend the ties with other CLARIN partners and to improve the interoperability and standard compliance of Slovak NLP resources and tools.