Slovak National Corpus – history and current situation

Since the second half of the 20th century we have witnessed the rapid development of the following disciplines, many of them being arbitrarily defined: sociolinguistics; psycholinguistics; pragmatic linguistics; text linguistics; cognitive linguistics. Particular positions are occupied by those disciplines that combine linguistic and mathematical methods, which began to develop with the introduction of cybernetics and with the interest in artificial intelligence, machine translations, etc. The increased performance of computer technology ushered in (and continues to bring about) new options for processing a large volume of data when processing natural language automatically. In the 1990s, large text corporawere being emphasised to such an extent that those years are titled the *corpus linguistics* decade. Besides the quantitative increase in the number of corpus workplaces and general national and specialised corpora (probably mainly in Eastern and Central Europe in the above decade), the early 90s were also characterised by a qualitative change in the attitudes of linguistics and other branches, and of interested experts, towards the corpus. A marked shift took place, from the question *"why corpus?"* to pragmatic considerations as to the best utilization of corpora, not only for improving the quality (increasing exactitude) or speeding up linguistic researches and their wider inter-disciplinarity, but also for the utilization of corpora as a reference source for information for various areas of science and research, as a tool for research and development of linguistic technologies and other application of artificial intelligence. The initial difficulties that characterised the introduction of corpora in the 1960s (inadequate output of computers, incompleteness of mathematical formalised descriptions of natural language and rejection by linguists, who were used to traditional theories that were based on small volumes of material that were often highly abstracted) were manifested in various forms in Slovakia thirty years later.

The establishment and operation of the Mathematical Linguistics and Phonetics Department of the Slovak Language Institute of the Slovak Academy of Sciences (today renamed as the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences) was the first promising development project in the area of mathematical and computer linguistics in Slovakia. Ján Horecký, who was its initiator and head, programmatically strove to develop the principles and methods of mathematical (algebraic) linguistics on the basis of the material of the Slovak language. Nevertheless, the lexicon of morphemes, which the department was preparing, was never finished.

In the next period, the mainly quantitative analysis of texts developed in the area of mathematical linguistics in Slovakia. J. Mistrík's frequency lexicons are well known, as well as the partial studies of some researchers who focused their attention on the statistics of linguistic phenomena.

Slovakia only subscribed to the worldwide trend of the development of computer and lin-

guistic technologies as late as 1989, when the topical area “Computer processing of lexis” was included in the programme of the symposium "Methods of research and description of lexis of the Slavonic languages”, which was held within the framework of the 7th Meeting of the Lexicology-Lexicographic Commission of the International Slavists Committee. It consisted of three Slovak (J. Horecký, J. Furdík, P. Žigo) and two foreign prepared contributions; 1 foreign and 1 domestic (J. Horecký) contribution to the discussion (cf. the proceedings of the homonymous symposium, 1990). V. Blanár glossed the topical area as follows in his closing speech: “The idea is being confirmed that the capacity of the human brain is not sufficient to master the continuing growth of information. Humans can meet many information and encyclopaedic challenges only with support from automatic data processing... Moreover, automatic data processing stimulates linguistic research... An important aspect is that such an approach requires looking at many linguistic phenomena from new points of view” (Blanár, 1990, p. 292).

More time passed between the verbalization and the implementation. It was characterised mainly by a lack of technologies and prepared experts in the area, but also by the steps that were directed systematically towards the establishment of the new discipline in Slovakia. After discussions on the options of co-operation of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences and the Information Centre of the Slovak Academy of Sciences, a new computer linguistics working group was established in 1990. The working group was headed by J. Horecký. They began to work on an integral concept of the future corpus of the Slovak language and lexical database (Jarošová, 1993). Work on a theoretical computer model of the Slovak language (Páleš, 1994) was an important element in this preparatory phase, but, the main work was a practical collection of texts in electronic form, and their first linguistic analyses (Benko, 1993; Šimková, 1993). The collection of the texts was extremely laborious due to the lack of technical and personal background; it was verbatim, word by word, without any tendency to representativeness or at least balance. An opportunistic approach was adopted, i.e., those texts were included in the corpus which were easily obtained and processed. No annotations were made (except for the basic bibliographic information) and the software equipment was also minimal (WordCruncher, later WordSmith; MicroConcord used to be used for preparation of concordances in the MS DOS mode).

The corpus of texts of the Slovak language was gradually made available up to 2002 for internal use within the framework of the Ľ. Štúr Linguistic Institute of the Slovak Academy of Sciences. In its final phase, the 30-million corpus included mainly journalistic texts, some texts of professional proceedings and journals, and a small quantity of belles lettres. A specific part of the corpus consisted of electronic versions of the following lexicographic productions of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences. Short Lexicon of the Slovak Language (issues 2 and 3); Rules of Slovak Orthography (1998); Synonymic Lexicon of the Slovak Language (issue 1); Academy Lexicon of Foreign Words; Lexicon of the Slovak Language (5 volumes).

One fact should be underlined, i.e., that even the minimal body of information available was in very active use from the beginning, for linguistic purposes (mainly lexicographic ones), and for the purposes of maintaining contacts with foreign corpus workplaces and projects. Most studies presented data processing technologies, selected statistical indicators, or foreign context and theory and practice of lexicographic utilisation of corpora, but there were also more lexical-grammar and comparison studies. The documentary material requested used to be individually prepared and provided to the authors of the above studies. Nevertheless, the existing corpus of texts of the Slovak language and the lexical database were the most widely used in the lexicographic team, which was preparing the concept of a big new monolingual dictionary of the Slovak language (its first volume is just about complete), as well as when preparing the 3rd and 4th issues of the Short Lexicon of the Slovak Language and the Rules of Slovak Orthography (issues 1998 and 2000). The knowledge and experience gained were honed in international events abroad and at home. The international seminar "Text Corpora and Multilingual Lexicography" was organised by Ľ. Štúr's Institute of Linguistics of the Slovak Academy of Sciences and Pedagogic Faculty of Comenius University in Bratislava in 1999. The event was organised within the framework of the international project TELRI II which took place within the framework of the European Commission programme INCO-COPERNICUS. The international seminar "Czech and Slovak Languages in Computer Processing" was organised by the same organisers in Bratislava in 2001 (the event with homonymous proceedings, 2001, resulted from participation in the above project).

This ad-hoc method for building and operating the corpus of texts of the Slovak language gradually showed itself to be impracticable in the long-term horizon. The most important aspect was that it was not comparable with the situation in the neighbouring countries. Moreover, demand for publicly accessible linguistic information began to increase in the late 1990s from the current users.. The demands of the lexicographers increased in the context of the volume and the structure of corpus texts, and the efficiency of their utilisation in conceptual work. More demands emerged within the context of Slovakia's accession to 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. The project assumed the establishment of a new specialized workplace with adequate technical and personnel background. Preparatory works were launched after the project was approved by the Government of the Slovak Republic on 13.2.2002. The works consisted of building and equipping workrooms in the loft of the building of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, and the purchase and installation of hardware and software. A working team of the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences was established at the end of 2002. The team comprises seven members.

Despite the fact that the corpus of texts of the Slovak language and lexical database had been built up in the Institute from ca 1993 to 2002, the Slovak National Corpus had no texts available, while contracts with providers of the existing texts either were not completed, or did not contain any clause that would enable incorporation of the texts into the corpus that would be accessible via Internet. Similarly, the technology for processing them did not comply with current standards. The corpus of Slovak language texts had been indexed (without any lemmatization and without any annotations, except for basic bibliographical data) and was operated under MS DOS by WordCruncher, which manifested marked capacity limits even at the level of 200,000 individual occurrences of words and at the overall capacity of 20 million words. The actual work on the building of the Slovak National Corpus (essentially from the beginning of 2003) was launched by the preparation of a licence agreement on other uses of the author's works according to the Authors Act, by the preparation of a concept of the structure of data in the corpus, and methods for their primary processing, i.e., conversion, tokenization, bibliographic and style-genre annotation (cf. Garabík, 2004; http://korpus.juls.savba.sk). In keeping with the tradition of the preceding corpus of texts of the Slovak language and in the context of other current projects, the Slovak National Corpus continues in part in its primary orientation to its user – the lexicographer. In addition, its scope was extended to the wider public (laymen interested in language, students, teachers, editors, and other persons who work with words and/or texts) and experts in the area of grammar research and in the area of NLP.

Our preparation of the project of the Slovak National Corpus was based on the following background: experience in the preparation of existing corpus projects, mainly in Czech; the requirements of potential users of the electronic database of Slovak texts; the real potential of the working group that is of a minimal size (a staff of seven persons), where persons from many different areas met, but which lacks graduates in computer or corpus linguistics, as no university has such disciplines on their curricula. The following basic objectives were listed in the concept of the Slovak National Corpus for 2003 – 2006 (Šimková, 2003, 2004):

1. Building a general monolingual corpus of written texts of the contemporary Slovak language (1955 – 2005) and making its representative part (200 mill. words) accessible via Internet; lemmatizing and morphologically annotating the accessible part; syntactically annotating a selected specimen.

2. Making the whole file of collected texts, which were electronically processed but bear no linguistic information, available to the staff of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, as well as to their external partners on the premises of the Institute, for the purposes of science and research, mainly for lexicographical purposes (the

scope is dependent on our technical background and on the willingness of our text providers).

3. Building specific corpora / databases

   – terminology database (in collaboration with the Ministry of Justice of the Slovak Republic and branch terminology committees);

   – database of lexicographical works (making available the lexicographical production of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences in electronic form via Internet, possibly on CD);

   – corpus of diachronic texts and corpus of dialect texts (on the basis of the needs of the researchers in the respective branches and according to technical background; mainly OCR of ancient prints or manuscripts and transcriptions of spoken language will be demanded);

   – parallel corpus/corpora (mainly for the so-called small languages, where such corporaare good tools for translators and interpreters, but also a good tool for making the language visible and accepted worldwide);

   – Corpus of spoken expressions (the technical and time demands for their transcription will require separate financial and personnel resources).

4. Creation of appropriate software tools (archiving texts; evidence database; conversions and filtrations of texts; lemmatizer; morphological annotator), use and adaptation of existing software tools (parser, corpus manager).

Our data collection was governed by the rule "as many texts as possible, as manifold as possible". Our approximation towards a representative sample of written texts in the current Slovak language was only very rough: one third consists of journalistic texts, another of fiction texts and the final third of specialized and non-fiction texts. Translations were prominent in the two latter groups, as they have a special position in small national and language societies (such as the Slovak one). Moreover, they were very poorly represented in the previous lexicographical manuals of the Slovak language. Approximately one third of translated fiction, specialized , or non-fiction texts were suggested for the Slovak National Corpus. 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.

Due to the acute need of materials for the team of lexicographers who were preparing the new monolingual dictionary of the current Slovak language, and conditioned by the accessibility and readiness of the provider of the texts, we agreed to accept any text in the first phase which could be gained without excessive effort (acquiring texts from approaching the provider through explaining the objective, the content, and the non-commercial character of the project, to the execution of the respective contract on the use of the work for scientific and research purposes in accordance with the law on copyright requires, on average, one or two months). In the next phase, we focused our attention on authors or publishers of specific texts which were missing in our representation of genres, or were not adequately represented (e.g. children's literature, the majority of specialized texts in the areas of natural and technical sciences).

When we had succeeded in concluding the starting number of contracts for the inclusion of texts into the corpus, we summarised the methodology of segmentation (tokenization) of Slovak text and its external, bibliographic and style-genre annotations. Concurrently, we initiated the preparation of the morphological tagset itself, as well as of the annotation tools (Forróová – Horák, 2004; Forróová – Garabík – Gianitsová – Horák – Šimková, in print). The texts gained were continuously processed and made available for use via the Internet. This approach could be demanding on users trying to become informed on the scope and structure of the texts that were effective at that moment. Nevertheless, the most important achievement was that they were able to work with Slovak texts. The first version prim 0.1 (primary, general corpus), made available in August 2003, contained 26 million tokens. The second version prim 0.2, made available in December 2003, contained 166 million tokens. The third version prim 1 with new tokenization and revised style-genre annotation, made available in July 2004, contained 192 million tokens. In the previous tokenization version, the final scope included paragraphs, titles, tags etc. As a result, version prim 0.2 actually contained fewer than 150 million tokens. Therefore, the increase between versions prim 0.2 and prim 1 was ca 50 million tokens. Moreover, version prim 1 was made available with lemmatization and also, internally, with morphological annotation that was implemented using the tagger and disambiguiser produced by the Mathematical and Physical Faculty of Charles University in Prague (authors J. Hajič and J. Hric). The current version, prim 2, was made available at the beginning of November 2005. It provides via the Internet to interested parties a corpus of 246 million tokens from almost 250 providers. In the context of licence agreements, the staff of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences have ca 10 million more tokens available (some providers of texts do not agree to the inclusion of their texts in the corpus on the Internet, but they agree to their availability for internal use within the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences, for instance, in the context of the preparation of the new monolingual dictionary). Besides the preceding automated morphological annotation, the latest version is also auto-

matically tagged on the basis of our own Slovak tag set (internal lexicographical annotation will be detailed in the next text).

The data structure of the Slovak National Corpus in the version prim 1 (that was the first version to provide a reasonable option of paying attention to style and genre classification) represented almost 182 million tokens (95%) from journalistic texts, 7 million from (3.5%) artistic texts, and 3 million (1.5%) from specialized and non-fiction texts. The disproportion in favour of journalistic texts was very marked. When presenting our corpus, we used to state that it was extremely unbalanced. Nevertheless, the share of non-journalistic texts was sufficiently relevant for us to create the first version of a balanced corpus *primvyv 1*. Within the framework of the basic structure with 60% journalistic texts, 20% fiction, and 20% specialized literature, it contained ca 12 million tokens. Balancing the entire range of corpus texts is also essential for the needs of morpho-syntactical research into the Slovak language in the corpus material (grant project Vega in collaboration with the Philosophical Faculty of Prešov University in Prešov). The project also investigates the distribution of language phenomena in specific types of texts. Representative selection of texts from the linguistics point of view can be influenced as a result, as well as for the purposes of other grammar researches on the corpus material. The frequency of tokens found in *primvyv 1* manifested the standard distribution not only of the most frequent prepositions, conjunctions, pronouns and particles, but also of the most frequent lexical words, as known from preceding researches and from analogical frequency findings, e.g. in the related Czech language, which were carried out on the representative corpus SYN2000 (Šimková, 2004).

The targeted collection of specific types and kinds of texts was clearly manifest in the new internal structuring of the current version *prim 2.0* as follows: 73% journalistic texts; 13% fiction; 4% specialized literature and non-fiction; 10% texts without the necessary annotation due to various reasons (work on its completion is ongoing). The proportion of translations into the Slovak language makes up 70% in fiction texts (more than 23 million out of 33 million tokens) and 46% in specialized texts (more than 5 million out of 11 million tokens). Our opinion is that this composition reflects relatively realistically the situation in the production and reception of the respective kinds of texts among Slovak readers, and it underlines the old querying of the orientation of the preceding excerption (prior to 1990) exclusively to top domestic production. The language of the translated texts is Slovak also, but is enriched by lexical and grammatical tools that also name other, unfamiliar facts and enrich the language system in this way. Due to the scope of the specialized texts (all of which were also included in the new balanced corpus, in such a way that their share is 20%, while some of the fiction texts, selected at random, were added so that their share makes up 20% and the remainder of journalistic texts, with the 60% share), the balanced corpus *prim 2.0-vyv* could be offered to the users of the Slovak National Corpus, with a volume of almost 56 million tokens.

Another important result of the new version of the corpus was an increased volume of texts dating from before 1990, resp. 1995, when no text existed in electronic form, or was not archived anywhere. Their share in the version *prim 2.0* is 17.5 million tokens. This could be attained only via intensive scanning and OCRing texts (almost 60,000 pages were processed per man-year in 2005) and their re-construction, which was performed in various volumes by ca 40 collaborators, mainly students. In the context of the goal of the project (i.e., to cover the thesaurus of the current Slovak language since 1955 and prepare material in this way, mainly for the purposes of conceptual works on the new monolingual dictionary of the current Slovak language), the investment is well substantiated. Nevertheless, there continues to be a marked lack of texts of specialized literature. Their representation in the corpus is necessary either in the context of the preparation of the above dictionary, or their use is planned in the context of the creation of a Slovak Terminological Database. The collection of texts (mainly those in the areas of technical and natural sciences) is obviously determined by the following factors: a) new scientific production in specific domains is more frequently written in foreign languages than in Slovak b) older scientific works are often considered obsolete and not relevant even from the point of view of terminology, and their authors are not disposed to make them available for any purposes.

After repeatedly mentioning the availability of the Slovak National Corpus for scientific and research and other non-commercial use via Internet, we should briefly mention the ways and options of working with it. First, searching in the Slovak National Corpus was implemented via a simple web interface (basic search without any support of regular expressions and without displaying external annotation). The corpus manager Manatee with the client Bonito (which was produced by the Faculty of Information Technology of Masaryk University in Brno, author P. Rychlý) could be used once contractual terms and conditions were agreed. The more recent versions of the Slovak National Corpus can be searched using our own corpus manager Korman, which was developed in the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences. The corpus manager facilitates the basic search including displaying bibliographic and style-genre annotation, as well as context extensibility. The corpus manager is available virtually for free: the searched string can be entered immediately on clicking agreement to the non-commercial use of the corpus on the introductory page. A specific form must be signed as the basis for using Manatee and Bonito. Then, the user gets his or her own password and has more statistic and frequency data available when searching the corpus as a whole or the studied expressions or forms. Average daily attendance on the corpus' web site is ca 200 entries. Ca 200 new users are registered annually. The individual password needs to be renewed by users at the beginning of each calendar year. This is a way to keep the database of users up-to-date, and discourage idle users. Foreign users are mostly from the neighbouring Czech Republic, but also those from Australia, Canada, Japan, Singapore, etc. can be found.

As previously mentioned, our work on the Slovak National Corpus up to the present also includes the share of the linguistics component. Nevertheless, due to its character, it is being built at a substantially slower pace, the first relevant results being obtained as late as in 2005. The rules of morphological annotation were in development from the beginning of 2003 Forróová – Horák, 2004; Forróová – Garabík – Gianitsová – Horák – Šimková, in press). They formed the subject of a discussion at the end of 2003 and, after minor adjustments, they were accepted as a basis of our own Slovak annotation. More differences emerged in automated morphological annotation when testing the tool that has been developed by the Faculty of Mathematics and Physics of Charles University in Prague: either in the approach of its authors (the so-called *engineering* approach, without any separation of some categories that are relevant for the Czech and Slovak languages, such as verbal aspect and incompleteness and the high error rate of the glossary of Slovak lemmas and forms), or in the language systems of the Czech and Slovak languages and in the theoretical assessment of some categories (e.g., adverbs, particles, secondary prepositions). The first phase of manual morphological annotation was launched at the beginning of 2004, using the co-operation of students of philological departments of other universities in Bratislava, Prešov, and Ružomberok. The tag set was gradually modified again (on the basis of our experience with the first annotations) and the annotation was also adjusted. G. Orwell's novel 1984 was annotated twice before the end of 2004, in the quantity of 102,000 tokens, and its corrections launched. In 2005 the corpus of the texts with manual morphological annotations was extended by the following selected texts with double annotations: the daily Sme and the Internet journal InZine (ca 50,000 tokens); non-fiction Internet encyclopaedia Wikipedia (ca 50, 000 tokens). The version *prim 2.0* was automatically tagged on the basis of the first version of the corpus with 130,000 tokens (manually annotated and corrected). Nevertheless, its error rate approaches ca 10%. Therefore, the phase of corrections of the results of the automatic morphological annotation was launched, in such a way that after manual annotation and disambiguation the corpus would have at least 1 million tokens and was adequate for training purposes for our own Slovak annotation tool. Developments also led to the launch of our own morphological analyser and generator of forms. The Slovak Dependency Treebank could be used for the purposes of improvement of the speed and efficiency of the corrections of morphological annotations. Work on the above corpus was launched in summer 2005 within the framework of the Slovak National Corpus, using technical tools and the linguistics and technical manual of the Faculty of Mathematics and Physics of Charles University in Prague. The first phase includes a double syntactic annotation of the texts that underwent manual morphological annotation. The next phase will include the option of linking manual morphological annotation on the analytical level and of automated morphological annotation.

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. Nevertheless, it is not a substi-

tute for orthographic or grammar manuals. It is only a basis for their creation, a basis that is readily accessible via Internet and essentially provides wider potential within the framework of the automated processing of large numbers of realistic texts. After completing its first big phase in 2006, its results should be made available on CDs/DVDs also. The next phase will include either a continuation of the work of building and balancing the primary national corpus and linguistic annotation of selected texts, or the work of the team and its partners will be oriented more towards the creation of the Slovak Terminological Database and building parallel corpora.

Bibliography

BENKO, Vladimír: Počítačové korpusy a analýza textu. In: Text a kontext. Zborník z medzinárodnej vedeckej konferencie. Text v priestore jazykovej komunikácie. Text v priestore literárnej komunikácie. Text v priestore didaktickej komunikácie. Prešov 18. – 19. novembra 1993. Red. F. Ruščák. Prešov: Pedagogická fakulta v Prešove Univerzity P. J. Šafárika v Košiciach 1993, s. 43 – 50. Martina

BLANÁR, Vincent: Na záver sympózia o metódach výskumu a opisu lexiky slovanských jazykov. In: Metódy výskumu a opisu lexiky slovanských jazykov. Materiály zo sympózia konaného v rámci 7. zasadnutia Lexikologicko-lexikografickej komisie pri Medzinárodnom komitéte slavistov (Nové Vozokany 24. – 26. apríla 1989). Zost. V. Blanár. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 1990, s. 289 – 292.

FORRÓOVÁ, Martina – HORÁK, Alexander: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 174 – 183.

FORRÓOVÁ, Martina – GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Návrh morfologického tagsetu SNK. In: Benko, V. (Ed.): Slovanské jazyky v počítačovom spracovaní / Computer Treatment of Slavonic Languages. VEDA, Bratislava 2005, s.

GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 164 – 173; aktuálna verzia http://korpus.juls.savba.sk.

GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. In:

http://korpus.juls.savba.sk, 2004.

JAROŠOVÁ, Alexandra: Korpus textov slovenského jazyka. In: Slovenská reč, 1993, roč. 58, č. 2, s. 89 – 95.

PÁLEŠ, Emil: SAPFO. Parafrázovač slovenčiny. Bratislava: Veda 1994. 305 s.

Slovenčina a čeština v počítačovom spracovaní. Ed. A. Jarošová. Bratislava: Veda 2001. 196 s.

ŠIMKOVÁ, Mária: Možnosti využitia programu WordCruncher pri analýze textu (na báze Sládkovičovej a Kraskovej poézie a ľudových rozprávok). In: Text a kontext. Zborník z medzinárodnej vedeckej konferencie. Text v priestore jazykovej komunikácie. Text v priestore literárnej komunikácie. Text v priestore didaktickej komunikácie. Prešov 18. – 19. novembra 1993. Red. F. Ruščák. Prešov: Pedagogická fakulta v Prešove Univerzity P. J. Šafárika v Košiciach 1993, s. 51 – 58.

ŠIMKOVÁ, Mária: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Zborník prednášok (Budmerice 22. – 23. máj 2003). Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.

ŠIMKOVÁ, MÁRIA: Slovenský národný korpus – východiská a plány. In: Slovenčina na začiatku 21. storočia. Ed. M. Imrichová. Prešov: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied 2004, s. 150 – 158.

http://korpus.juls.savba.sk