Slovak National Corpus

Alexander Horák, Lucia Gianitsová, Mária Šimková, Martin Šmotlák, and Radovan Garabík

Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences Bratislava, Slovakia
korpus@juls.savba.sk,
http://korpus.juls.savba.sk/

Príspevok odznel na konferencii „TSD 2004“ (Brno 2004) a bol publikovaný v zborníku.

Abstract. Presented paper informs about the current status of Slovak National Corpus, goals and achievements of the project, as well as about technical details concerning the building of representative, balanced, annotated corpus of modern Slovak language.

1 History of Corpus

1.1 Prehistory

Corpus linguistics has relatively weak tradition and history in Slovakia, although the first steps were taken already during 1962 – 1970, with creating of the Department of Mathematical Linguistics and Phonetics of the Slovak Language Institute[1], lead by J. Horecký. The department presented only some partial results, mostly concerning statistical analysis of various aspects of Slovak language. Revival of mathematical linguistics came in 1989[1], but the proposed field of interest was oriented mostly towards computer aided lexicography. It was not until 1991 when the first idea[2] of creating computer database of Slovak language appeared, including lexical database as well as representative corpus of written language.

1.2 Text Corpus of Slovak Language

What followed was the creation of so-called Text Corpus of Slovak Language[3], which was just a set of ad-hoc collected texts, without any guidelines concerning representativeness or coverage. The corpus had been used internally in Ľ. Štúr Institute of Linguistics, mostly for lexicographic purposes.

1.3 Slovak National Corpus

Text Corpus of Slovak Language continued its growth until 2002, when it became clear that there was a need for a new version of corpus that would be representative, annotated and available to professional community. On 13 February 2002

[1] now Ľudovít Štúr Institute of Linguistics

the Government of Slovak Republic approved the Slovak National Corpus project and a project to computerise linguistic research in Slovakia. Since that time, Slovak National Corpus took a systematic approach in creating desired corpus of Slovak language texts.[4]

2 Design Criteria and Principles

When we speak about Slovak National Corpus, we mean actually two different things. The first is Slovak National Corpus as a department of Ľudovít Štúr Institute of Linguistics, taking part in several different projects, among them the text corpus of Slovak language, also having (the corpus project) the name Slovak National Corpus. However, it should be sufficiently clear from context if we are speaking about the department or about the corpus.

Our current plans consider building a text corpus consisting of texts published during 1955–2005, the lower limit is dictated by a substantial spelling reform carried out in 1953 (and giving two years to "settle down" the spelling). We expect less texts to span most of this time interval, carefully choosing texts to maintain uniform time and genre distribution, while the rest of the corpus will cover more recent dates, when electronic versions of texts are readily available. Desired size of the corpus is 200 million words, which we assume to collect during 2005. At the time of writing, the corpus contains 187 million words, but is highly unbalanced, consisting mostly of journalistic texts.

As a minimal requirement by potential users, the corpus should be lemmatized and contain information about morphology, as well as bibliographic annotation. We expect to create a "core", consisting of manually lemmatized and morphologically annotated texts of about 1 million words, that could be used to train morphological analysers and other NLP tools to be used with the rest of the corpus.

Access to the whole corpus (excluding texts with too prohibitive copyright licenses) is publicly available free of charge on internet, in exchange for a simple registration procedure.

3 Data Structure

Corpus is divided into 4-level hierarchy, consisting of following parts:

- Archive – this is were original texts are kept, in original formats, without any conversion or modification.
- Bank – here we keep texts converted into common text format, used as the source for any following analysis/transformation, together with bibliographic and style/genre annotation.
- Corpusoid[2] – texts on this level are lemmatized and morphologically annotated, and any other eventual additional linguistic information should be added here.
- Data – binary form of data as used by corpus manager.

[2] for the lack of better name

3.1 Data Format

Bank contains texts in our own internal XML-like format, trying to keep as much as reasonably achievable amount of non-textual information of the original files (e.g. information about presence of images, tables, typographic settings etc.). Using proper XML here is not reasonable, because of its rather strict requirements concerning tag hierarchy and structure, quite unlike the chaotic flow of typographical and other additional information present in source data.

On the other hand, corpusoid contains texts in XCES format[5]. We are aware of the fact that XML format has (from technical point of view) many shortcomings and pitfalls and is not really suitable for large databases of linguistic data, but it is a standard, there are many tools dealing with it, and XCES (or variant thereof) is used by other corpora as well, so the exchange of data and tools becomes viable.

3.2 Data Flow

After obtaining the source data and permission to use them, the files are placed into the Archive for further analysis. For each file format, we have to use a specific tool to convert it into our intermediate format for the Bank. There is often some manual work involved, especially in extracting texts from proprietary file formats, using non scriptable software.

After the extraction, the texts are analysed and conversion tools to unify typographic conventions are written (unifying opening/closing/straight quotes, different types of dashes, paragraph and line separators, ways of text emphasis, removing hyphenation). If possible, texts are automatically split into smaller units – documents (e.g. articles, in the case of newspaper and journals), and each document is assigned automatically generated bibliographic and genre/style annotation – it is often possible to apply some heuristics based on known source medium layout to find out more about the document (e.g. TV schedule starts always on the same page, first page is always devoted to headlines, sport news are titled accordingly etc.). When converting fiction and other larger texts, these are annotated by human annotators, using simple WWW interface build over the Bank.

From the Bank, texts are converted into XCES format. The most difficult obstacle we encountered was in writing tools to make a well formed hierarchical XML out of original flat texts, intermingled with typographic markup. For example, `<hi>` tag denotes emphasis, and can occur either as part of one paragraph, or can span over several paragraphs. In the former case, we want `<hi>` to appear as a subnode of `<p>` node, while in the latter, we have several `<p>` subnodes of one `<hi>` node. Recognising the correct hierarchy of tags is the most complex part in our software utilities converting the texts from the Bank into XCES format.

Subsequently, the documents in the Corpusoid are tokenized, morphology analysers are run on them, and the final lemmatized, morphosyntactically annotated texts are written back into the Corpusoid. Texts from the Corpusoid are then converted into vertical format, suitable for input into the corpus manager.

4 Bibliographic and Genre/Style Annotation

We keep all the usual bibliographic information, if available, such as author, date, source. As for style/genre annotation, we keep three-way system, distinguishing genre (such as poem, song, letter, manual, novel, short story...), type (art, informative, scientific), and domain (life style, law, politics, natural science, technical texts...). There is also finer division into subgenre/subtype/subdomain categories, if applicable.

5 Lemmatization and POS-tagging

5.1 Existing Tools

There exists one working tool for morphological annotation of Slovak language, based on similar tool for Czech language[6,7]. There is also promising development of Slovak version of AJKA[8], We are deploying the former analyser, while we are also actively looking into the possibility of using AJKA once it is ready, simultaneously extending each tagger with additional data from both of them. As for morphologic tagset, both these analysers are using their own versions.

5.2 Morphological tagset

When designing morphological tagset for Slovak language, we had two main sources of inspiration: the first was positional system as used in [6], the second one uses key-value pairs and is used by AJKA. Both approaches have their advantages and disadvantages. While conceptually being equal, we have to keep in mind that given tagset is going to be actively used for information retrieval by humans, so we have to try to make the tagset as palatable as possible. Above mentioned positional system is lean and compact, but tries to squeeze different number of grammatical categories into given number of positions, sometimes with awkward results. The second system has more logic in it, but tags are long and confusing. We tried to make our own system, that has advantages of both these approaches, while trying to keep out the disadvantages. The basic principle lies in dividing tags into categories, each category is identified by paradigm[3] and POS type of given word. Each category then has its own corresponding set of marks, each mark being represented by one alphanumeric character, specifying one grammatical category. We tried hard not to use the same characters for different categories, even if applying to different paradigm/POS type classes. The result is tagset that is simultaneously positional (in scope of one paradigm/POS type) and each character denotes unique grammar category.

The complete set of categories and marks has been described elsewhere[9], we give here just some short examples – e.g. "ženou" has tag "SSfs7" which should be read as:

[3] or to be precise, by corresponding set of similar paradigms

S	substantive
S	substantive paradigm
f	feminine gender
s	singular
7	instrumental case

whereas preposition “v” is tagged with “Eu6” that should be read as:

E	preposition
u	non-vocalised
6	binds with locative case

6 Corpus Manager and User Interface

Slovak National Corpus currently uses Manatee[10] as its prime corpus manager. There is also a simplified WWW interface to our own corpus manager, that we are actively working upon.

7 Other Projects

Slovak National Corpus takes part directly in several additional projects, among them it is the initiator of Slovak WordNet project and the new comprehensive Slovak Language Grammar based on corpus research (together with Prešov University). The plans to build parallel Slovak-Russian corpus (together with Department of Mathematical Linguistics of St. Petersburg University), parallel Slovak-Croatian corpus (together with Croatian National Corpus) are underway.

References

1. 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. April 1989).* Edited by V. Blanár. Bratislava: Jazykovedný ústav Ľ. Štúra SAV (1990)
2. Benko, V.: Slovak Language Lexical Database. In: *Computational Lexicography.* Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences (1991) 5–8
3. Jarošová, A.: Korpus textov slovenského jazyka. In: *Slovenská reč 2* (1993) 89–95
4. Šimková, M.: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. Budmerice: Počítačová podpora prekladu (2003)
5. Ide, N., Bonhome, P., Romary, L., XCES: An XML-based Encoding Standard for Linguistic Corpora. In: *Proceedings of the Second International Language Resources and Evaluation conference.* Paris: European Language Resources Association (2000)
6. Hajič, J., Hladká, B.: Czech Language Processing - POS Tagging. In: *Proceedings of the First International Conference on Language Resources and Evaluation.* Granada, Spain: (1998) 931–936

7. Hajič, J., Hric, J., Kuboň, V.: Machine Translation of Very Close Languages. In: *Proceedings of the ANLP 2000.* Seattle, U.S.A. (2000) 7–12
8. Sedláček, R.: Morfologický analyzátor češtiny. PhD. thesis. Faculty of Informatics, Masaryk University Brno, (1999)
9. Forróová, M., Garabík, R., Gianitsová, L., Horák, A., Šimková, M.: Návrh morfologického tagsetu SNK. Bratislava: Slovko 2003 (2003)
10. http://www.textforge.cz/