Style-Genre Annotation of the Slovak National Corpus

Alexander Horák†

22nd June 2004

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

Abstract

The aim of the article is to show the first results of the style-genre annotation process of Slovak National Corpus and the current state of the text distribution in Slovak National Corpus according to linguistically relevant attributes of the style-genre annotation schema.

1 Introduction

The project of Slovak National Corpus has started in year 2002 as continuation of the so-called Text Corpus of Slovak Language previously built at the L. Štúr Institute of Linguistics in the years 1990 – 2001[?]. This corpus was characterized mostly by its rather small size (about 20 million tokens in 2001) and highly unbalanced state - vast majority of the texts were from the journalistic domain, at the same time containing an unomittable amount of lexicographic texts like the Short Diconary of Slovak Language (1998), Comprehensive Dictionary of Slovak Language (1954 – 1969) as well as linguistic periodicals like Linguistic Review (Jazykovédný časopis), Slovak Word (Slovenská reč) and Word Culture (Kultúra slova). On one hand the fact of the higher distribution of journalistic texts was could be positively taken from the lexicographic point of view (neologisms etc.), but on the other it was no use for larger stylistic or any other statistically based lexical researchers when taking into account the higher percentage of the linguistic texts which are characteristic by metatext word usage distorting the "normal" (meaning typical) usage of the language. Moreover, there was no clearly defined conception of the corpus data architecture and storage of the texts constituting the corpus. The corpus was available as raw
text chunked into pieces according to some attribute that they had in common (author, title, volume etc.). Each of the chunks written as a separate file was searchable with the WordCruncher or WordSmith software tools. Unfortunately technical limits (especially in WordCruncher) did not allow for a more explicit linguistic markup of the texts on the level of document, not speaking about the level of the document’s logical structure, or even word’s morphological information (lemma, tag). It is obvious that under these circumstances this, so to say, opportunistic corpus could not qualify as a fully-reliable source of linguistic information that could be used in serious research (it was, though...).

2 The Annotation and Its Implementation in Slovak National Corpus

Having the experiences and a user feedback from the previous attempt to build a corpus of slovak texts it was quite clear from the very beginning of the Slovak National Corpus project that a new conception of the corpus data structure as well as its markup must be worked out. A large pressure was put on its transparency, clear-cutness and system-independency. These features could only provide better interchangeability and reusability in the future. As an important starting point for we considered the 7 corpus maxims proposed by G. Leech[?]:

1. Possibility to extract original corpus
2. Possibility to separate annotations
3. Based on well defined guidelines
4. Make clear how the annotations were done
5. Make clear that there may be errors in the corpus
6. Widely agreed theory-neutral annotation scheme
7. No annotation scheme is the a priori standard scheme

In following, we will try to argue that all the maxims mentioned above were kept in our approach.

2.1 The Slovak National Corpus Data Structure

First, we give the structure of the Slovak National Corpus database. It is designed as a four-level hierarchy[?]:

- Archive – this is where original texts are kept, in original formats (doc, pdf, pm, rtf, html and others), without any conversion or modification.

- Bank – here we keep texts converted into common text XML-like format, used as the source for any following analysis/transformation, together with bibliographic and style/genre annotation.

- Corpusoid – this level is generated from the Bank. Except bibliographic and style-genre annotation texts are lemmatized and morphologically annotated, and any other eventual additional linguistic information should be added here. The texts are stored in XCES format.

- Data – binary form of data automatically generated from the Corpusoid as used by corpus manager.

Figure 2: Position of a file and its annotation on the level of Archive.

//net/jano/data/snk/archive/2003/02/28/a/

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Size</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>..</td>
<td>Directory</td>
<td></td>
<td>2003.02.28 12:15:08</td>
</tr>
<tr>
<td>.annotation</td>
<td>Directory</td>
<td></td>
<td>2003.04.23 10:51:36</td>
</tr>
<tr>
<td>Tutorial_sk.lyx</td>
<td>LYN</td>
<td>97 KB</td>
<td>2003.02.28 12:04:08</td>
</tr>
</tbody>
</table>

As could be seen from the above the data hierarchy allows easy and separate access to original corpus texts as well as the annotation, which is the
confirmation of the first two maxims. On both Archive and Bank levels, the annotation is stored in a separate file, whereas on the level Corpusoid the XCES format allows keeping text and its markup in a single file also having the annotation in a separate file.

2.2 Information in the Annotation Schema

There has been a lot of discussion concerning the question of quality and quantity of the metainformation on the corpus data. On one hand there are informations which are objective and not dependent on the linguistic interpretation of the text (like author, publisher, date), on the other hand we have a large variety of theories giving us a detailed network of fine-grained categories according to which text could be interpreted. When speaking about the linguistic annotation of texts in general we refer to concepts of annotation previously formulated e.g. by Geoffrey Leech[?] or joint-research international initiatives like TEI[?] or EAGLES[?]. The annotation schema was largely inspired by EAGLES recommendations and experiences of Czech National Corpus.

References

Figure 5: A Corpusoid file in XCES format showing the header of a document and few tokens from its body.


