

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 proces 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 Ľ. 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 Dicitonary of Slovak Language (1998), Comprehensive Dictionary of Slovak Language (1954 – 1969) as well as linguistic periodicals like Linguistic Review (Jazykovedný 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 chararecteristic by metatext word usage distorting the "normal" (meanining 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

Figure 1: Main WordCruncher corpus manager window.



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/`

Name	Type	Size	Time
 .	Directory		2003.04.14 17:31:32
 ..	Directory		2003.02.28 12:15:08
 .annotation		1 KB	2003.04.23 10:51:36
 Tutorial_sk.lyx	LYX	97 KB	2003.02.28 12:04:08

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

Figure 3: Position of a file and its annotation on the level of Bank.

`file:///net/jano/data/snk/bank/2003/12/12/e/1/`

Name	Type	Size	Time
.	Directory		2004.03.24 13:03:28
..	Directory		2004.03.24 13:03:16
.annotation		1 KB	2004.06.04 15:13:08
vsetko sluzi na dobre.bank	BANK	188 KB	2004.03.24 13:02:46

Figure 4: Position of a file and its annotation on the level of Corpusoid.

`file:///net/jano/data/snk/corpusoid/2003/12/12/a/1/`

Name	Type	Size	Time
.	Directory		2004.06.10 18:03:10
..	Directory		2004.06.10 18:02:52
.annotation		1 KB	2004.06.10 18:03:10
corpusoid.xces	XCES	1,249 KB	2004.06.10 18:03:10

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 Informations in the Annotation Schema

There has been a lot of discussion concerning the question of quality and quantity of the meta-information 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

- [1] Garabík, R.: Slovak National Corpus. *Unpublished manuscript.*

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

```

<TEI.2>
<teiHeader status="new" type="text">
<fileDesc>
<snkAnotation>
  <Bogocong>MDan1</Bogocong>
  <ISBN>80-88949-61-0</ISBN>
  <Author>Mária Danieliszová - Jesenícká</Author>
  <Corrected>tru</Corrected>
  <Origlang>YYY</Origlang>
  <Emphasis>tru</Emphasis>
  <Bibliography>TTT</Bibliography>
  <Subdomain>soc</Subdomain>
  <Type>img</Type>
  <Varieta>std</Varieta>
  <Anotator>Helena</Anotator>
  <Conglomerate>Mária Danieliszová - Jesenícká - Všetko slúži na dobré</Conglomerate>
  <Medium>lib</Medium>
  <Origname>YYY</Origname>
  <Sourceid>2003-12-12-e</Sourceid>
  <Transsex>YYY</Transsex>
  <ISSN>YYY</ISSN>
  <Rhyme>nrh</Rhyme>
  <Dateorig>TTT</Dateorig>
  <Translator>YYY</Translator>
  <Diacritics>tru</Diacritics>
  <Genre>mem</Genre>
  <Translation>org</Translation>
  <Subgenre>bel</Subgenre>
  <Lang>slk</Lang>
  <Name>Všetko slúži na dobré</Name>
  <Subtype>pro</Subtype>
  <Origauthor>YYY</Origauthor>
  <Comment>dopisat DATE, DATEORIG, BIBLIOGRAFIU!!!</Comment>
  <Domain>hum</Domain>
  <Date>TTT</Date>
  <Paragraphs>tru</Paragraphs>
  <Authsex>fem</Authsex>
  <Id>2003-12-12-e-1</Id>
</snkAnotation>
</fileDesc>
<titleStmnt>
  <title>Všetko slúži na dobré</title>
  <author>Mária Danieliszová - Jesenícká</author>
</titleStmnt>
</teiHeader>
<body>
<h1>
<p>
<s>
<hi>
<tok>
<orth>Mária</orth>
<disamb>
<base>Mária</base>
<ctag>NNFS1----A----</ctag>
<hrest>;Y</hrest>
</disamb>
<lex>
<base>Mária</base>
<ctag>NNFS1----A----</ctag>
<hrest>;Y</hrest>
</lex>
</tok>

```

- [2] Leech, G.: Corpus Linguistics. The Current State of Art. In: *Acta Universitatis Carolinae. Philologica*.
- [3] Tadić, M.: Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. In: *Filologija 30-31, Zagreb, 337-347*.
- [4] Tadić, M.: Napredak u radu na Hrvatskom nacionalnom korpusu. *Drugi hrvatski slavistički kongres, Osijek 14-18. rujna 1999*.
- [5] Text Encoding Initiative: The Standard for the Corpus Annotation. In: *Proceedings of the Second International Language Resources and Evaluation conference*. Paris: European Language Resources Association (2000)
- [6] Expert Advisory Group on Language Engineering Standards: Preliminary Recommendations on Text Typology. <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>
- [7] Český národní korpus. <http://ucnk.ff.cuni.cz/>