Old Sources and Modern Procedures: Computer Processing of Old-Church Slavonic

Kiril Ribarov

Research fellow
Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranske nam. 25, Prague 1
ribarov@ckl.mff.cuni.cz

Abstract

A framework for computer processing of Old-Church Slavonic including its specific features is presented. The corpus of Old-Church Slavonic and its annotation is introduced. Incorporation of manually pre-prepared card catalogues into a corpus is proposed.

1. Introduction

An inevitable part of the Slavonic cultural and linguistic heritage is the area of paleoslavistics, in which, during the past years an acknowledgeable piece of work has been done towards corpus and formal processing of the material of the common dead language of the Slavs, the Old-Church Slavonic language (OCSL).

Computer processing of OCSL was given official acknowledgement at the First International Conference for Computer Processing of Medieval Manuscripts, held in Blagoevgrad, Bulgaria, in 1995. On this conference, the first of its kind, it was clearly shown that the various isolated but similar efforts can be unified into a stream of common tendencies leading into computer oriented approaches to the study of OCSL material. On the basis of our own direct experience (Ribarova, Ribarov 1995), (Ribarov, Ribarova 1998), (Ribarova, Ribarov 1998) and (Camuglia G. et al. in print) we would like to present in this paper the corpus processing of the OCSL material. As a sequence of continuous work, the following sub-areas of interest are considered:

- Lexical annotation of a corpus of OCSL, its creation, management, access and distribution;
- Methods, tools and procedures for corpus processing of OCSL related to the above stated aspects;
- Incorporation of manually created card catalogues within a corpus.

This work, together with works of similar kind as (Camuglia 1996), (Paskaleva, Dobreva 1995), and (Miltenova 1995) form the state-of-the-art of the computer processing of Slavonic medieval manuscripts and early printed books.

2. The OCSL Corpus

The biggest collection of annotated Old-Church Slavonic (OCS) manuscripts (Table 1) forms what is referred to as the OCSL Corpus (of the Macedonian Redaction).

Manuscript Name[1]	Time of creation	Size: Number of folios
Bitolský (Kièevský) triod	11th/12th century	101
Boloòskýž altáø	1230-1241	264
Grigorovièùv parimejník	12th/13th century	104
Chludovùv triod	end of 13th century	191
Krninský damaskin	end of 16th century	315
Lesnovská pareneze	1353	315
Macedonské záhøebské evangelium (Mihanoviæovo evangelium)	14th/15th century	146
Orbelský triod	second half of 13th century	245
Pogodinùvž altáø	13th century	278
Stanislavov prolog	1330	321
Vatašski minej	1453	228

Table 1: OCSL Corpus Contents

The eleven sources form a collection of 2508 folios of text of almost 2 million word forms[2]. The manuscripts are electronically processed and annotated in the Institute for the Macedonian Language, Skopje, Macedonia, using the STINO ver.1 (Ribarova, Ribarov 1995) software package (its new version is to be presented in section 4).

2.1. The Structure of the Corpus Material

Using the STIN-O-SANCT procedures (Section 4.1) each manuscript can be exported in an SGML-like structure. The structure of the manuscript data is displayed here.

2.1.1. The Manuscript Header

[1] The names are listed as originally inserted in the corpus database (in Czech).

[2] Although there are efforts to standardize the OCS characters they are still not a part of the standardized set of characters. One of the most detailed proposals for their standardization can be found in (Birnbaum 1995).

The header description is stated in the sequel (The "*" symbol shows where multiple entries are possible).

```
<datasource T>  //T is  manuscript ID
 <header>
  <zd=T> //T is manuscript abbreviation
  <name=T> //T is the manuscript name
  <flag=N> // internal, additional information
  <created=T> //T is the date and time of creation
  <anr=T> //T is redaction abbreviation
  <ant=T> //T is translation abbreviation
  <ost=T> //T determines the manuscript sorting type[3]
  <tst=T> //T determines the translation sorting type
  <time=T> //T is time flag
  <mn=T> //T is lower time ID
  <mx=T> //T is upper time  ID
  <mud=T> // T is locality information
    <location=T> //T is the place of location of the original of the manuscript
  <tr=T> //T is typization of the manuscript
  <fonttv=T> //T is font name used for comments
  <fontzl=T> //T is font name of the manuscript
  <fontp=T> //T is font name of the translation
  <fonto=T> //T is font name for the other fields
  <note>
      // place to enter any kind of notes related to the manuscript
  </note>
 </header>  // end of the header information
 <f>
      // this tag defines a word form entry, see subsection 2.1.2 The Manuscript Contents.
 </f>*
</datasource>
```

2.1.2. The Manuscript Contents

In the following the SGML-like structure of each word form is presented. The "*" symbol is used to mark the possibility of repetition of the entry. Implicitly, this description reveals also the rich annotation of the manuscripts. The manuscripts in the current version of the corpus do not posses all of the information as presented below. They miss the redaction keyword distinction, which was not supported by the software during their annotation. Further, the translation is a translation of the manuscript from an older source (mainly used for parallel/critical linguistic study, and e.g. reconstruction and restoration of the damaged part of the manuscript) and does not necessarily exist for a manuscript or can exist only for some of its parts (mainly translations from a Bible written earlier in Ancient Greek); therefore the translation is not a target value, but a source for the processed words of the manuscripts. The older version of the annotation software did not support assignment of keywords to a translation, definition of collocation spans and discontinuous complexes.

The structure of each word form is:

```
<f>
 <g>S {S       original form}
  <a=N1 f=N2C l=N3 p=N4>
      {N1      unique position in the document}
      {N2      folio number}
      {C       position on the folio}
      {N3      line number}
      {N4      position on the line}
<r>S          {S       rendered form}
  <t= S1 c= S2>
          {S1      correlation type}
          {S2      correlation }
  <x=C N1 N2>
          {C       complex type}
          {N1      connection of the rendered
  form to a form to the left}
          {N2      connection of the rendered
  form to a form to the right}
  <e N1 N2>
          {N1      left environment collocation
  size in number of words}
          {N2      right environment collocation
  size in number of words}
  <s>
   <l=S1 g=S2 h=S3>
          {S1      lemma}
          {S2      POS}
          {S3      sense disambiguation}
   <v=S>*         {S       keyword paradigm}
   <r=S1 l=S2 g=S3 h=S4>*
          {S1      redaction type; abbreviated
  string for redaction}
          {S2      lemma}
          {S3      POS}
          {S4      sense disambiguation}
  </s>*
  <p>S    {S       translation of the form}
   <v=S>  {S       translation specification,
meaning}
    <l=S1 g=S2 h=S3>
          {S1      lemma}
          {S2      POS}
          {S3      sense disamb iguation}
  </p>*
  <m=S>* {S       morphological specification of
the rendered form}
 </r>*
</f>
```

2.2. The Importance of Being Old

The presence of OCSL within the framework of formal processing of Slavonic languages, is of great importance mainly from the following points of view:

- Synchronic aspects: completeness, in a broader sense, of various phenomena found in OCSL (on various levels), thus enriching the global view on linguistic phenomena within the Slavonic languages;

- Diachronic aspects: processing a language being a complex system evolving in time, benefits from the experience from the work on the OCSL material forms solid bases for both, theoretical and computational aspects, which are to occur, sooner or later, from its conceptual point of view within any framework of computational processing of a current language;

Furthermore, if a corpus of a language is there in order to witness the language, then designing a corpus of a dead language is more than tempting, since the corpus can

[3] The OCS documents may require various sorting orders depending on the fact whether the manuscript has been originally written in Glagolitic or in Cyrillic. Various (di/multi)graphs should also be supported.

include all the available sources, and thus can become a real and complete witness of that language.

3. Some specifics of the OCS Corpus

Although a computer linguist may consider that the problem of codification is a resolved one, the work on an old manuscript brings again the question of character representation and its digitization. The main difficulties are due to the facts that: the text is not written as a linear stream of characters, the text can be written in a continuous way, and characters not known in advance can suddenly appear in the text. Even a glyph that can be considered as a variant of a known character does not necessarily have to be a variant. The latter fact makes a standard codification almost impossible; each scholar justifies his needs of using various variants of specific characters in a different way[4]. Therefore, one should distinguish two aims, which are related to the problem of codification:

- Rewriting a manuscript in order to restore it and reproduce the original;
- Rewriting a manuscript in order to extract and preserve its lexical content[5].

A linear variant of a certain group of characters[6] may have non-objective influences depending on the knowledge and available sources of the scholar who renders the manuscript. Various revisions and different renderings (by other scholars) may lead to different conclusions about the (real) linguistic content of the considered group of characters.

Therefore a word form W is defined to be the pair W = (*a*, A), where *a* is a group of characters (a picture of A, represented as graphics or text, without the need of parsing or editing), while A is a set of linear renderings of *a*. In order a manuscript to be annotated, the cardinality of A must be at least one for each *a*, and *a* must be present. Let $a_L \in$ A, be a linear rendering (rendered form) of *a*; a_L[7] is always accompanied by *a*.

Examples of variability in rendering are shown in Table 2.

Original Form: *a*	Possible renderings: A
с҃	с‹ти›ˣ
с͡х	с‹ти›х с‹ти›х‹иρа› с‹ти›х‹єρа›
ѡ҃ц҃а	ѡᵀ‹ь›ца ѡᵀца

Table 2: Examples of some variabilities in rendering

For the rendering phase a special set of symbols is developed[8]. Table 3 summarizes their meaning and usage.

Symbol	Explanation	Example
/	The word continues on the next line.	.....глагол/ ати.....
//	The line continues.	..... ρєчє г‹оспод›ь// своимъ оучєникомъ
]	Missing end of the word	прѣвѣзꙋ·ѥ· → прѣвѣзꙋ]носите прѣвѣзꙋ]
‹ ›	Abbreviated form	бг̀ → б‹ог›ъ
[ ]	Conjecture: Substraction Adding	момои → [мо]мои ρади → ρа[сѫ]ди!
( )	Reconstruction (partial visibility)	шбл(ѣчє) (са)
(..)	Damaged part of a document	бъскрнч(..)
–	Haplography	памятвоѭ → памя–т т–воѭ
!	Error	прагъ! (=circulus)
_	Set phrase	бъ_слѣдъ
+	Composite form (complex)	крстили+ са+ ѥсте
<*n*>+ or +<*n*>	<*n*> is any number within a collocation determining a relative number of words to the left/right the current word should be associated with	крстили+2 1+са+1 1+ѥс те
»,	Change of location Verse	», or »Is, etc. Arabic number in the text
, ,	Foglio	,1r, or ,1v, etc.
// //	Number or continuos phrase (This symbol might be accompanied by other markers for its further sub-classification.)	//ρкꙁ// //оч·г·итє//

Table 3: Signs for rendering of forms

Textual analysis of each word form is not only a technical task, since OCS manuscripts contain features that inevitably force a scholar to take some linguistic decisions and to solve problems at a syntactical level in order to be able to reach the semantic level and to face its problematic and incomprehensible parts as well.

Those features are[9]:

- scriptum continuum,
- resolving of variants at various levels of the language,
- abbreviations,
- rendering of damaged and unknown parts,
- correlation to other sources,
- lemmatization and lemma disambiguation.

These features also directly influence the processing of the manuscripts and the design of the data structures.

[4] These problems are very similar to those that one would experience if trying to digitize/codify hand written notes with various abbreviations, sub and super scripts, figures, arrows etc.

[5] Lexical content as understood by the scholar(s); through the lexical content a wider linguistic content is considered.

[6] This group of characters from the original manuscript - if isolated and rendered -should represent a word form.

[7] In the linguistic works only a_L has so far been considered a word form. This has led to the development of various and more or less successful (but never unambiguous) norms of rendering.

[8] Where possible these symbols follow the traditional ones used for manual rendering.

[9] For more details see (Camuglia G. et al. in print).

Apart from large variability, multiple entries even after disambiguation e.g. for keywords should also be allowed/considered since the context is the only available disambiguation "tool".

4. Software tools

Although there are many corpus oriented software products nowadays, it is still difficult to select the proper one in general, and even more difficult if one would like to start with corpus processing of a language being not formally preprocessed and/or a language for which there is not at least a small corpus available. In order to fill this gap, and to allow a quick and consistent initiation of such a process, a special software package was designed: STIN-O-SANCT (Charles University, Prague, Czech Republic). This software package is specially designed for OCSL corpus processing. Besides that, its interface is flexible such that it can be easily accustomed in order to serve as a database of a new corpus for both, flective and analytic contemporary languages.

The specific needs of OCSL corpus processing, and the direct experience of OCSL corpus processing collected during the work for the past ten years allowed us to incorporate within STIN-O-SANCT, various flexible modules and to consider various phenomena, including the particular (but preserving the common) ones within a language system. Some of those are: extreme variability present at all levels of the language system (each word form can have more than one reading, each of those readings can have: different contextual behavior, various morphological roles, various lemmatizatoins, various translations), multi-redaction normalization, cross-reference correlation to other sources.

4.1. STIN-O-SANCT

STINO (Staroslovenski Textovi INdeksiranje i Obrabotka) (Ribarova, Ribarov, 1995) is the original pilot software package by which the OCSL Corpus word forms have been processed and annotated (Ribarov, Ribarova, 1998; Ribarova, Ribarov, 1998). Recently, STIN-O-SANCT[10] (with the haplographic part O-SANCT: Old-church Slavonic ANotated Corpora of Texts) has been designed to serve as a central software package for OCSL corpus processing.[11] STIN-O-SANCT is an Internet application totally programmed in Java with a single database server.[12] The client uses the Internet protocol to communicate with the server. In the sequel we would like to point out the concept of data structures, and the possibilities of annotation offered by this program.

The central element is the a_L from W. Each a_L (more than one possible variant supported) is associated with its original form a. Each a_L can be annotated so as to:

- be the member of a dynamic collocation;
- be the member of a complex (continuous or discontinuous; including analytic forms, phrases, idioms);
- have more than one keyword (redaction specific keywords and identificational keywords);
- have more than one translation equivalent;
- have very rich morphological tags;
- additional cross references (e.g. Biblical information);
- unique ID from the text citation.

By a keyword we understand: a lemma, disambiguation sign for the lemma, POS information, and a list of semantic/paradigmatic characteristics.

Two kinds of keywords are used: an identificational one and a redaction-specific one. A keyword specified within a certain redaction (and characteristic for it) is the redaction-specific keyword, the only keyword used so far in classical works. The identificational keyword is redaction independent and could be defined as the member of a coverage of disjunct sets of all of the redactions. Thus mapping is formed between any two redactions (through the identificational keyword). In general, the relations between the redactional keywords are neither one-to-one nor simple. They can be of any type: one-to-one, one-to-many, many-to-one or many-to-many. The situation can be even more complicated since it is possible to have the case in which a part of the keyword corresponds to a part of another keyword within another redaction. Very often, there is no relation between some keywords of different redactions. The identificational keyword is proposed to serve as a bridge among redaction-specific keywords allowing for a OCSL general dictionary.

A translation equivalent is a translated word form and its keyword, associated to W and its identificational keyword.

There can be more than one identificational keyword associated with a word form and more translation equivalents associated with each identificational keyword.

The environment supports (ready document exports from the system in Rich Text Format) multi-redaction keywords, generation of various types of indices including index verborum, retrograde indices on both word forms and keywords, indices of translation equivalents, redaction-dependent indices, collocational dictionaries.

Various multi-criteria searches are supported within a single manuscript and within any subset of manuscripts.

The program is user-friendly with automated administration, thus easy to use.

To speed the linguistic work and to assure homogeneity during the annotation, a history-based annotation of renderings, lemmatization, morphological tag assignment and translation equivalents was implemented.

The input file can be a rendered or a non-rendered manuscript file in plain text, Rich Text Format, SGML-like format or the format can be defined by the user. The input word units are parsed according to the characters from Table 3 and inserted into a document/manuscript database.

[10] A software project headed by Kiril Ribarov, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

[11] A software package designed for the Laboratory for Computer Processing of OCSL, Slavonic Institute of the Czech Academy of Sciences, Prague.

[12] For a detailed description of this program we refer to the STIN-O-SANCT Manual Pages.

5. Card catalogues

One of the latest issues being considered is inclusion of a great amount of manually processed lexicographic material, which has a form of a card catalogue, into a corpus. The card catalogues are usually treated as 'dead' and static sources of the language. But thanks to the context included for each word form and thanks to the unique identification of the word from within a manuscript it is possible to 'dynamize' this kind of information and to design a way of a faster digitization of it.

A typical card of the card catalogue incorporates the following information:

- lemma
- additional lemma (serves for more specific definition of the lemma, usually in multi word components)
- word form
- morphological identification of the word form
- word form ID from the manuscript
- correlation of the word form to other sources
- context of the word form
- translation of the word form, including the context of the translated part.

The contexts are large enough as to connect one to each other - being the key observation for the card catalogue-to-corpus insertion. Currently intelligent context connections are under development.

This characteristics allows:

- faster insertion of the cards into a specially designed database, and
- manuscript from card catalogue reconstruction.

It is the manuscript to card catalogue reconstruction being of very big importance for otherwise impossible check on the correctness of the manual word from extraction.

Procedures for computer assisted insertion of the contents of each card are designed in order to minimize the manual part of the work based on the very frequent repetition of the same word form (from different locations) over the card catalogue (texts). Also, check mechanisms can be provided to insure the consistency and reliability of the material.

The card catalogue-to-corpus insertion is currently under implementation.

6. References

D. Birnbaum (1995), Informational and presentational units in early Cyrillic writing, First International Conference on Computer Processing of Medieval Slavic Manuscripts, 24-28 July, Blagoevgrad, Bulgaria.

Camuglia M. (1996), The Psalter, its tradition and the computer: a new method of textual analysis, «Palaeobulgarica», XX, 1.

Camuglia G., Camuglia M., Ribarov K. (in print), To appear in internal publication edited by A. Zampolli, Instituto di Linguistica Computazionale CNR, Italy.

Miltenova Anisava (1995), Computer Assisted Analysis of the Description of Slavic Manuscripts. Proceedings of the First International Conference on Computer Processing of Medieval Slavic Manuscripts, Blagoevgrad, Bulgaria, pp. 129-139.

Paskaleva E. and Dobreva M. (1995), New tools for old language: computer processing of Old Bulgarian texts, First International Conference on Computer Processing of Medieval Slavic Manuscripts, 24-28 July, Blagoevgrad, Bulgaria.

Ribarova Z., Ribarov K. (1995), Computer Processing of Old-Church Slavonic, Results and Prospects, in Proceedings of the First International Conference on Computer Processing of Medieval Slavonic Manuscripts, Blagoevgrad-Princeton, Bulgaria.

Ribarov K., Ribarova Z. (1998) A Time for a Corpus of Old Church Slavonic, in Proceedings of the Workshop on Varieties in Slavonic Texts, Sofia.

Ribarova Z., Ribarov K. (1998), Living Conservation of the Lexis of Old Church Slavonic, «Palaeobulgarica», XXII, 2.