TEI encoding of Serbian proverbs

Cvetana Krstev¹, Duško Vitas², Vesna Šatev³

cvetana@poincare.matf.bg.ac.yu, vitas@poincare.matf.bg.ac.yu, vesnasatev@ptt.yu,

¹Faculty of Philology, University of Belgrade
²Faculty of Mathematics, University of Belgrade
³Faculty of Agriculture, University of Belgrade

Keywords: TEI encoding, Serbian Language, Language Resources, Serbian Proverbs

Abstract

In this article we discuss some problems in the computer processing of Serbian text, as well as possible solutions for resolving them. One solution is illustrated by the production of the electronic edition of one classical text, the collection of Serbian proverbs assembled by Serbian language reformer Vuk Stefanović Karadžić. This text is marked using the TEI (Text Encoding Initiative) P5 standard, visualized as HTML file, and searched by extensive index.

1. Introduction

Serbian language belongs to the family of Slavonic languages. It has some characteristics which have to be considered during text processing. Some of them are:

a) **Rich inflective and derivational morphology**, which results in many different word forms associated to one lemma.

b) **Usage of different variants in both spoken and written texts**, for instance *mleko* and *mljeko* ‘milk’, in Ekavian and Ijekavian pronunciation respectively.

c) **Regular usage of two alphabets**, Cyrillic and Latin.

All these characteristics influence the ways Serbian texts are prepared and processed. Only the good initial preparation of a text can enable its subsequent effective processing.

2. The classical edition of Serbian proverbs

The collection of proverbs has been assembled by Serbian language reformer Vuk Stefanović Karadžić. It contains besides proverbs also frozen expressions, curses, oaths and other common phrases. Its first edition dates from the year 1849. It comprises of 6919 proverbs and phrases. All the later editions, including the edition published jointly by NOLIT and PROSVETA in 1987, reflect this first edition in many aspects: for instance, the elements of the old orthography are present (*neću* instead of *neću*, ‘I will not’) as well as the characters that do not exist in modern Serbian Cyrillic alphabet (for instance, the hard sign ћ). The inventory of proverbs has not changed either, although there is number of references to the proverbs that are actually not listed as well as a certain number of proverbs that do not differ but for the word ordering (for instance, the proverb number 4286 *Po dvaput se u vodenici govori* and the proverb number 5748 *U vodenici se po dvaput govori* ‘Twice is spoken in the mill’) or minor lexical differences (for instance, the proverb number 3763 *Ni luk jeo, ni lukom vonjao* and the proverb 3764 *Ni luk jeo, ni njim mirisao* ‘Niether has eaten the onion, nor is stinking onit’).
In 1989 the once distinguished Belgrade publishing house NOLIT started the project of modernization of this collection that ended in 1996 with the publication of the new edition. Except for the adjustment to the modern orthography and the correction of typographical errors, the original text was not otherwise altered: it contains the same repository of proverbs and phrases that are separated historically in four parts.

<table>
<thead>
<tr>
<th>GNIZDO</th>
<th>F= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2049</td>
<td>Kakovo gnizdo takva tica,/ kakav otac takva dica.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNJEZDA</th>
<th>F= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5040</td>
<td>Soko iz vranjega gnijeza.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNIJEZDA</th>
<th>F= 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6259</td>
<td>Što je veća tica, veče joj gnijezdo treba.</td>
</tr>
<tr>
<td>4641</td>
<td>4641 Rdava je tica, koja u svoje gnijezdo tori.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNJEZDU</th>
<th>F= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4746</td>
<td>Svaka je kvočka na svome gniježdu jaka.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNJAVI</th>
<th>F= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0706</td>
<td>Gnjavi mačiče.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GNJAT</th>
<th>F= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>D177</td>
<td>I na vrat i na gnjat.</td>
</tr>
</tbody>
</table>

Table 1. An excerpt from the concordances of the original text; in the second column are lemmas added by lexicographers.

The text is accompanied by the comprehensive index that contains all the lexical words occurring in the text of proverbs: nouns, verbs, adjectives and numbers. The first step in the index compilation was the production of concordances that served as a basis for the lemmatization of all nouns, verbs, adjectives, and numerals (Table 1). The second step was performed independently by two experienced lexicographers that followed traditional lexicographic principles. Their choice of lemmas, however, differed in many cases due to graphic variations not properly covered in existing dictionaries (Table 2). Many of these variant forms are in use even today and can be both spoken and written.

<table>
<thead>
<tr>
<th>NB</th>
<th>word form</th>
<th>1st lemmatizer</th>
<th>2nd lemmatizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2956</td>
<td>bregovi</td>
<td>breg</td>
<td>brijeg</td>
</tr>
<tr>
<td>5589</td>
<td>brijeg</td>
<td>brijeg (=breg)</td>
<td>brijeg</td>
</tr>
<tr>
<td>295</td>
<td>brijegu</td>
<td>brijeg (=breg)</td>
<td>brijeg</td>
</tr>
</tbody>
</table>

Table 2. What is the lemma for a word form bregovi, breg or brijeg?

As a result, all variations used in the source text where reproduced in printed index and are linked as references of a type see also. A certain number of keywords not actually found in the text were introduced in order to make the index more suitable for the modern user: they are the source of the references of a type see. The presence of all the variations encumbered significantly the index. The index contains a number of archaic terms as well, and for them some short explanation is given. For instance, some entries from the printed index that include the examples from Table 2 are given in Figure 1.

The index itself was produced automatically from the concordance files that contained the information added by lexicographers. This had several advantages, among others it was possible to identify the proverbs that contain only functional words – for instance the proverb number 192 (see Appendix) consists only of the auxiliary verbs and conjunctions. In the whole collection there are 82 such proverbs, some of which are quite long, for instance Што је твоје оно је и моје и твоје, а што је моје оно је само моје ‘What is yours it is both yours and mine, and what is mine it is mine only’. It is maybe interesting to add that the index
was not only automatically produced but also automatically formatted by incorporating the
plain TeX commands; however, the publisher was not capable or willing at that moment to
use the index formatted in this way in its final edition.

Figure 1. The structure of the produced index and its presentation in the new paper edition.

3. The electronic edition of Serbian proverbs

Electronic edition of Serbian proverbs was based on this last NOLIT’s paper edition, and it
was marked according to the TEI standard\textsuperscript{1}. In order to do the proper encoding, the structure
of the text was analyzed and the elements that were going to be marked recognized. A few
examples from Figure 2 illustrate the typical structure of a proverb or a saying in this
collection. Every collection item is given a unique identification number. It consists of a text
of a proverb or other type of phrase, which is followed by some explanation that is not
obligatory (e.g., see the proverb number 4518). This explanation can sometimes include the
references to other items in the same collection (e.g. see the proverbs number 1167 and 1233).
The saying sometimes has the form of a short song (e.g. the proverb number 1233). Both
saying and the explanation are usually very short, but they can occasionally consist of more
then one paragraph.

169 Беспослени поп и јариће крсти. Беспослен чоек забававља се свакојаким беспослицама.
192 Било па (и) није. Било па прошло, и већ нико и не мисли да буде друкчије него што је.
419 Боље (ти) је да те ћера Турчин са сабљом него Швабо са пером. У Хрватској.
1167 Ђе двоје диште треће се пише. Рече се за мужа и жену. Гледај! Ђе се
dвоје ваљају трећему се надају.
1233 Ђе се двоје ваљају трећему се надају. Гледај: Ђе двоје диште треће се пише.

\textsuperscript{1} First the SGML version P3 was used; later it was adapted to the XML P3 version.
Since the work on the electronic edition started in the beginning of the 90s when Unicode was still developing, and Cyrillic alphabet was not well supported either by font or by software developers, and having in mind that Latin alphabet was in Serbia then as well as today in equal use, the whole text was encoded using the Latin alphabet and plain ASCII code, with specific letters of Serbian Latin alphabet represented by appropriate entity references. The list of these entities is given in Table 3. The usage of these entities enables the text encoded in this way to be transformed unambiguously in both Latin and Cyrillic alphabet.

<table>
<thead>
<tr>
<th>Entity reference</th>
<th>Replacement Text</th>
<th>lower case Latin/Cyrillic</th>
<th>Entity reference</th>
<th>Replacement Text</th>
<th>Upper case Latin/Cyrillic</th>
</tr>
</thead>
<tbody>
<tr>
<td>&amp;zx;</td>
<td>ž</td>
<td>џ, ж</td>
<td>&amp;ZX;</td>
<td>Ž</td>
<td>З, Ж</td>
</tr>
<tr>
<td>&amp;sx;</td>
<td>š</td>
<td>š, ш</td>
<td>&amp;SX;</td>
<td>Š</td>
<td>С, Щ</td>
</tr>
<tr>
<td>&amp;cx;</td>
<td>ć</td>
<td>ц, х</td>
<td>&amp;CX;</td>
<td>Ć</td>
<td>Ц, Љ</td>
</tr>
<tr>
<td>&amp;cy;</td>
<td>č</td>
<td>ћ, љ</td>
<td>&amp;CY;</td>
<td>Č</td>
<td>Џ, Љ</td>
</tr>
<tr>
<td>&amp;dx</td>
<td>đ</td>
<td>д, ђ</td>
<td>&amp;DX;</td>
<td>Ď</td>
<td>Д, Ђ</td>
</tr>
<tr>
<td>&amp;lx</td>
<td>&quot;lj&quot;</td>
<td>лj,љ</td>
<td>&amp;LX;</td>
<td>&quot;Lj&quot;</td>
<td>Лj,Љ</td>
</tr>
<tr>
<td>&amp;nx</td>
<td>&quot;nj&quot;</td>
<td>нj,њ</td>
<td>&amp;NX;</td>
<td>&quot;Nj&quot;</td>
<td>Нj,Њ</td>
</tr>
<tr>
<td>&amp;dy</td>
<td>&quot;dž&quot;</td>
<td>dž,џ</td>
<td>&amp;DY;</td>
<td>&quot;ДŽ&quot;</td>
<td>Дž, Ј</td>
</tr>
</tbody>
</table>

Table 3. List of entities used to enable both Latin and Cyrillic alphabet representation

At the first moment it looked as if the structure of the text is rather specific, so some experiments were done to expand the TEI tag set with some specific tags (Krstev 1997; Vitas 1999). Later it was decided to stick to the standard TEI tag set and to accommodate the specific structure features of this collection by use of the appropriate values of the element attributes. Moreover, only the elements of the core tag set were used. In order to adapt to the TEI set of tags, it seemed most appropriate to use the <div> element both for the collection item as a whole, and for its main constituent parts: text of a proverb and its explanation, while they are distinguished by the attribute type. The global attribute id is used for the item unique identification. The cross-references are established by <ptr> element. The value of its attribute resp identifies who established the reference, since not all of them were present in the original text (e.g. see the proverb number 169 in the Figure 2 and in the corresponding encoding). The encoding of the same illustrative items from Figure 2 is given in Appendix.

Many phrasal elements are marked as well. First of all, all typographic conventions from the original text are marked; for instance, the embraced word in the item 419 from Figure 2 is marked with the tag <hi>. The value of the attribute rend points the type of highlighting; the value PA suggests that the element content is in parentheses, which usually means that the content is optional. Since electronic edition was produced on the bases of the chosen paper edition, all the errors found in it were corrected and duly marked (the use of the element <corr> in the collection item 192). In quite a number of sayings the obscene, or rather not very polite, words were used, and they were usually given in the abbreviated form. For such cases the <abbr> tag is used with expan attribute whose value is the full form of the word (see item 4518).
In the proverbs themselves, as well as in the text of the explanation many proper names appear, such as geographic, personal and ethnic names, as well as dates. They are all tagged as an `<name>` element, with the attribute `type` appropriately set. For instance, in the proverb number 418 the ethnic name “Turk” and the place name “Croatia” appear, while in the proverb 1264 the dates designed by appropriate saints appear.

The element `<note>` is also quite frequently used. Sometimes it is used for the explanatory notes that are interlaced with the proverb text, such as directives as how some proverb should be spoken. More often it is used to tag the text segment that can be used instead of some other segment that precedes it (see item 3488 in Figure 2).

The electronic edition is supplied by TEI header, in which only the obligatory elements are present. The production of electronic edition was never part of some official project and was never financially supported, so there was no spare time to prepare and maintain it properly. So the extensive changes of electronic edition that occurred during more than ten years of its existence and development are unfortunately not properly recorded.

4. The index for electronic edition

The index in the electronic edition is organized as a list, using the element `<list>`, in which each `<item>` is dedicated to one index entry. Each index entry was automatically assigned an identification number that became the value of the general attribute ID of the element `<item>`. The keyword itself is represented by the element `<term>`, which can occasionally include some explanation in the element `<gloss>`. The part of speech of a keyword is represented by the element `<note>`. Besides this, one or more lists can be nested into the element `<item>`, due to the fact that `<list>` is an interlevel element and the content of `<item>` element is defined as `specialPara`. Those nested lists can perform various functions:

- the identification number of proverbs containing the keyword are listed (list type is `ProverbsID`),
- the identification numbers of index entries are listed that represent the synonyms or pronunciation variants of the keyword (list type `SeeAlsoRef`), in which case the entry contains also the list of a type `ProverbsID`, or
- the index entry is actually a dummy entry since it does not contain the list of type `ProverbsID` but it contains the identification numbers of index entries that do contain the actual references (list type `SeeRef`).

The structure of the index is illustrated by the same examples given in Figure 1:

```xml
<item n="ID2">
  <term>abenjak</term>
  <gloss>kapa od abe; </gloss>
  <note>i</note>
  <list type="ProverbsID">
    <item>5713</item>
  </list>
  <list type="SeeAlsoRef">
    <item>ID1620</item>
    <item>ID1918</item>
  </list>
</item>

<item n="ID266">
  <term>breg</term>
  <note>i</note>
  <list type="SeeRef"/>
</item>
```
5. The visualisation and searching

The text of Serbian proverbs and index of terms that occur in them, prepared in previously described way, provides a various number of ways for visual representation and for further processing. As one of them, we chose HTML representation by using XSLT (Extensible Stylesheet Language Transformation). HTML documents are produced for both text of proverbs and index. Those documents are supplied with hypertext links connecting related proverbs or term-proverb pairs. Hypertext links are also produced for term-term pairs for those terms that occur in <list type="SeeRef"> or <list type="SeeAlsoRef"> elements. In that way the effective and user-friendly text search is obtained. The part of the.xslt file which establishes the hypertextual link between a term and all the proverbs in which it occurs is:

```xml
<xsl:for-each select="list[@type='ProverbsID']">
  
  <font size="-1">Poslovice u kojima se pojavljuje:</font>
  <xsl:for-each select="item">
    
    <font color="#FFFFFF">  
      <a href="E-VUK.xml#{text()}"><xsl:value-of select="text()"/></a>
    
    </font>

  </xsl:for-each>

</xsl:for-each>
```

Therefore, the user can browse through the index and easily retrieve all the proverbs related with any of the terms from it. Similarly, terms related to each other by „See“ or „SeeAlso“ references can also be followed easily.

The Figure 3 shows the Web representation of the index. The Web representation of proverbs is formatted in the similar way. Although prepared as a web presentation it is not publicly available. During the period of transition, the publishing house NOLIT faced the bankruptcy, then it was privatized and today it still trying to find its way to the passed glory. Thus, at this moment it was not possible to discuss with them the copyright issues.
6. Future development

We do not think that the work on the electronic edition of proverbs is finished, since only now that the text has been properly encoded, so that its visualization and searching can effectively be done, many new possibilities are opened. For instance the proper typology of all items has to be done in order to distinguish proverbs from other types of phrases.

The searching of the collection can also be improved. All the data, lexical resources and tools are available for the automatic establishments of hypertextual links that go from the proverb to index thus enabling the user who browses through the text of collection, to click on any word of his/her interest that would lead him directly to the corresponding index entry, from where he/she can continue the search in the already established way.

One line of future development is particularly appealing. Having in mind that this text dates from 1849 and that many terms in index are archaic, it is interesting to establish which of these proverbs are still in use, where the lexical changes have occurred, and which of them are really frozen. In order to do this we plan to use the corpus of contemporary Serbian (Vitas et al. 1999), and also the Serbian portion of the Web.

The preliminary results show that many of the proverbs are still in use. For instance, for the proverb number 169 the following usage was found in the corpus of contemporary Serbian:

Gospodin Dindić liči na besposlenog popa koji krsti jariće, uzalud se trudi da ostavi utisak kako radi važan posao.
Mister Dindić looks like an idle priest that baptizes the lambs, he is in vain trying to make the impression that he is doing the important work.

This example shows that lexica of the proverb have not changed; it has, however, been syntactically adapted for the particular purpose. Let’s consider the proverb number 368:

Боље је кукавицу у руци но сокола у планини (имати).
Better (have) cuckoo in a hand then a falcon on a branch.
The corpus search, as well as a search on a web shows, that although the proverb is still widely in use the lexica has changed, and today cuckoo and falcon are rarely mentioned being replaced most often by sparrow and pigeon. The proverb is so well known that is often used in shortened form, which easily conveys the intended meaning:

Dok se bivši bugarski car Simeon Drugi premišlja hoće li biti premijer (iz carske perspektive vrabac je već u ruci) čitav niz monarha u egzilu, i izdaleka, sanja svoj kraljevski san o "velikom povratku".

While ex-Bulgarian tsar Simeon II considers whether he is going to be a prime minister (from his tsar's perspective sparrow is already in his hand), the sequence of monarchs in exile dreams from afar a royal dream of a "big comeback".

Also, according to the context, both sparrow and pigeon can be replaced in order to achieve some specific effect, as suggested by the following occurrence from the Web:

Vrijedi princip: “bolje informacija 95% točnosti u ruci nego ona od 100% na grani”.

The principle is valid: “better information 95% accurate in hand then the one 100% on the branch”.

All these examples have been manually obtained. We are planning in the future to elaborate a tool that would use the text of proverbs, produced index and electronic dictionary of Serbain (Vitas et al. 2003) in order to automatically formulate the queries and search the Web in the quest of the potential usage of the proverbs.

References


Krstev, C. 1997. “Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije” (One approach to text modeling and transformation), PhD Thesis, Faculty of Mathematics, University of Belgrade


Appendix

&lt;div type="provSection" id="P169" n="169"&gt;
&lt;!-- The idle priest baptizes even lambs --&gt;
    &lt;div type="proverb"&gt;
        &lt;p&gt;Besposlen pop i janča krsti.&lt;/p&gt;
Besposlen &eek zabav&lx;a se svakojim besposlicama. <ptr target="P168" resp="CV"></p>
</div>
</div>
</div>
</div>
</div><div type="provSection" id="P192" n="192">
<!-- It was, so it isn't -->
<div type="proverb">
<p>Bilo, pa <hi rend="PA">i</hi> nije.</p>
</div>
<div type="explanation">
<p>Bilo pa pro&sx;lo, i ve&cx; ni ne <corr sic="masli" resp="CV">misli</corr> da bude druk&cy;i je nego &sx;to je.</p>
</div>
</div>
</div><div type="provSection" id="P419" n="419">
<!-- It's better for you to be chased by a Turk with a saber then by a Germen with a pen.-->
<div type="proverb">
<p>Bo&lx;e <hi rend="PA BD">ti</hi> je da te &cx;era <name type="ethnic">Tur&amp;cy;in</name> sa sab&amp;lx;om nego &name>&SX;vabo</name> s perom.</p>
</div>
<div type="explanation">
<p>U &name type="place">Hrvatskoj</name>.</p>
</div>
</div>
</div><div type="provSection" id="P1168" n="1167">
<!-- Where two breathe the third is written -->
<div type="proverb">
<p>&DX;e dvoje di&amp;sx;e tre&amp;cx;e se pi&amp;sx;e.</p>
</div>
<div type="explanation">
<p>Re&amp;cy;e se za mu&amp;zx;a i &zx;enu. Gledaj! <ptr target="P1234" resp="Vuk"/&DX;e se dvoje va&amp;lx;aju tre&amp;cx;emu se nadaju.</p>
</div>
</div>
</div><div type="provSection" id="P1234" n="1233">
<!-- Where two are rolling for third they are hoping -->
<div type="proverb">
<lg type="song">
 &DX;e dvoje va&amp;lx;aju/
 &DG;e se dvoje va&amp;lx;aju
</lg>
</div>
</div>
</div>
</div><div type="provSection" id="P1264" n="1263">
<!-- On the St George's day the outlaws meet, on the St Mitar's day the outlaws separate -->
<div type="proverb">
<lg type="song">
 &DX;ur&amp;dx;ev danak&lt;/name>
hajdu&amp;cy;ki sastanak,&lt;/p&gt;
 &DX;e dvoje di&amp;sx;e tre&amp;cx;e se pi&amp;sx;e.&lt;/p&gt;
</lg>
</div>
</div>
</div>
Eq. 1. It’s not possible that the wolves are fed and all the sheeps (lambs) are present →

<note type="replacement">kozlići</note> na broju.

U primorju i u Crnoj Gori govori se i kozlići mjesto ovce.

Eq. 2. He would speak by his a.s (if his mouth were tied) →

(da mu usta sve u)