

# Digitization of the Serbian folk proverbs compiled by Vuk S. Karadžić

This paper aims to present the digitization process of a very important piece of Serbian intangible cultural heritage, Српске народне пословице и друге различне као оне у обичај узете ријечи (Engl. Serbian folk proverbs), compiled by Vuk Stefanović Karadžić during the first half of the 19th century. In the paper, we discuss the necessary steps in the digitization process, the challenges we had to deal with as well as the solutions we came up with. The goal of this process is to have a fully digitized, user-friendly version of Serbian folk proverbs, that will also easily integrate and be compatible with other digitized resources and/or multi-dictionary portals.

CCS Concepts: • **Applied computing** → *Extensible Markup Language (XML)*; *Arts and humanities*; **Digital libraries and archives**; **Annotation**.

Additional Key Words and Phrases: Annotation, TEI, Electronic editions, Cultural heritage, Serbian language

## ACM Reference Format:

. 2021. Digitization of the Serbian folk proverbs compiled by Vuk S. Karadžić. 1, 1 (October 2021), 11 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Српске народне пословице (Engl. Serbian folk proverbs) [Stefanović Karadžić 1965] is a vast collection of Serbian folk proverbs, compiled by Serbian linguist and philologist Vuk Stefanović Karadžić (1787-1864) during the first half of the 19th century, starting as early as 1814 [Pantić 1965]. Vuk Stefanović Karadžić was one of the most important reformers of modern Serbian language and he played a major role in the process of modernization of the Serbian language in the 19th century. His work went beyond the language reform, as he was a meticulous collector of Serbian folklore, including oral epic and lyric songs, folk tales, proverbs, etc. He is considered, by many, to be the first Serbian folklorist, ethnographer, and literary critic [Deretić 2004]. He drew the attention of European scholars and readers to Serbian folk poetry and Serbian culture as a whole, due to his acquaintances with the times' leading scholars, such as Wolfgang von Goethe, Jacob Grimm, and many others. Proverbs were seen by Vuk as an important component of the Serbian folk tradition and, in addition, as a perfect representation of the language he advocated for to later become the new standard [Čupić 1981]. He dedicated a considerable amount of time during his travels to collect as many proverbs as he could. The result was two editions of Serbian folk proverbs – the first one was published in Cetinje (Montenegro) in 1836, containing more than 4,000 proverbs, and the second, the extended one, in Vienna in 1849, with over 6,000 proverbs.

Proverbs collected by Vuk in Serbian folk proverbs are very diverse, and they cover a vast number of topics, providing insight not just into everyday life, but into the core of what Serbian folklore is. Furthermore, Serbian folk proverbs contain more than just proverbs – Vuk also included curses, oaths, and frozen expressions. Since they were perceived by Vuk as the most authentic expression of the language of its speakers, most of them were kept in their original form,

---

Author's address:

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2021 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

1

often with more than one variant of a single proverb. Therefore, it is obvious why Serbian folk proverbs hold such importance in the realm of Serbian intangible cultural heritage, and why the decision to digitize them was a logical next step.

There were some previous attempts to digitize Serbian folk proverbs [Krstev et al. 2006]. However, the results obtained from this project are not visible to the general public, so we have no knowledge of the outcome and whether the digitization included the complete work of Serbian folk proverbs and to what extent they were digitized.<sup>1</sup>

The need for the digitization of Serbian folk proverbs is obvious, given the significance it has for Serbian culture, especially in this age where the preservation of intangible cultural heritage is a given. Still, some other factors contribute to the need of having a fully digitized and user-friendly edition of Serbian folk proverbs.

As far as the authors of this article know, Serbian folk proverbs are currently available for the general (and scientific) public as a printed edition and as an e-book (PDF file) on Google Books. When deciding to digitize Serbian folk proverbs, the first step was to see into detail what the faults of its printed edition are, and how to resolve them in the digital one. First and foremost, printed editions are final. Once out of the print, there are no more opportunities to add, fix, or replace anything. On the other hand, digitization is an ongoing process that leaves us with more options to upgrade and change any inconsistencies or errors we may encounter, especially with older editions, such as Serbian folk proverbs. Another factor that makes it much more difficult to navigate through the material is the material itself – a strict lexicographic approach (including alphabetical order) may not be the best solution when dealing with proverbs, since in most cases there is more than one variant of a saying. Digitizing the material from Serbian folk proverbs would make it easier to connect the variants, and to connect the proverbs on several other criteria – meaning, usage, semantic field, etc.

Bearing all the pros and cons of the printed edition in mind, as well as characteristics of the digital edition we would like to obtain, we set up the following steps for the digitization process of the Serbian folk proverbs.

## 2 THE DIGITIZATION PROCESS

In the process of digitization, we relied on the experience of previous projects performed within our institution and followed their workflow. Along with “four types of methods and activities for creating digital representations of lexical resources: 1) image capture; 2) text capture; 3) (lexicographic) data modeling and 4) (lexicographic) data enrichment” [Tasovac and Petrović 2015], which we would refer to here as phases of digitization, we added one more, 5) data presentation.

According to [Tasovac and Petrović 2015], the phrase image capture pertains to “the process of recording the visual representation of the text by means of digital cameras and scanners and its subsequent delivery to the user as a digital image”. In our case, this activity has already been done by the Serbian National Library, so we used the digital object they had created.<sup>2</sup>

The next step was to obtain the textual content of the proverbs in digital form. Following [Tasovac and Petrović 2015], we have also named this phase text capture, when referring to the “transposition of textual content into a sequence of alphanumeric characters”. This task was also executed by the Serbian National Library, automatically, by means of Optical Character Recognition (OCR) software, which converts images into searchable strings. More about this phase will be presented in chapter 2.1.

Data modeling refers to “the process of explicitly encoding the structural hierarchies and the scope of particular textual components” [Tasovac and Petrović 2015]. This implies marking up both the macrostructure and the microstructure of

<sup>1</sup>See also: <http://alas.matf.bg.ac.rs/~cvetana/proverb/#deo1> (accessed on October 25th 2021).

<sup>2</sup>Accessible at: [https://digitalna.nb.rs/attach/NBS/Tematske\\_kolekcije/procvat\\_pismenosti/sabrana\\_dela\\_vuka\\_karadzica/II\\_146423\\_09/output.pdf](https://digitalna.nb.rs/attach/NBS/Tematske_kolekcije/procvat_pismenosti/sabrana_dela_vuka_karadzica/II_146423_09/output.pdf)

the text. Marking up the macrostructure pertains to the demarcation of the smallest independent units of the texts (such as entries of the dictionaries) while marking up the microstructure involves further segmentation of each unit (in the case of the lexicographic data, differentiation of lemmas, grammatical information, senses, etc. inside each entry). This kind of text annotation dramatically increases the use-value of the lexical resource, enabling retrieval of more reliable and faster search results, for instance, in the case of proverbs collection, retrieving all instances of a particular lexeme when it appears only in the text of the proverbs and not in the text of the explanation. For further information on the data modeling phase see chapter 2.2.

The next phase, data enrichment, involves “the process of encoding additional information that specifies, extends or improves upon the information already present in the lexicographic resource” [Tasovac and Petrović 2015]. This kind of text annotation can also increase the use-value of the lexical resource, enabling, inter alia, multiple access paths to the information from the resource. For instance, if each proverb was annotated with the semantic field that would mark its domains of language use, the user could see a list of proverbs only from one domain, from which point he could easily access and further explore individual proverbs from the list. Such access to the information could facilitate research on certain historical or ethnolinguistic topics as well as enable the user to retrieve a specific proverb, overcoming the obstacle of not knowing the exact form of the proverb. For more information on the data enrichment phase in our project see chapter 2.3.

## 2.1 Text capture

As previously mentioned, we worked on the text already extracted by the Serbian National Library using Optical Character Recognition (OCR) software. Although OCR output of Cyrillic texts, especially with older versions of characters, usually contains more errors than texts in Latin, this one was very accurate.

The correction of the OCR output was done manually by a group of fifteen high school students as a part of the seminar on digital humanities during the 2021 summer school of the Serbian language in Tršić. In addition to correcting errors made by the program, typographical errors found in the printed edition were also corrected. The following editorial decisions were aimed to preserve the original characteristics of the printed edition:

- Preservation of the old orthography: in the printed edition we used as a model, the text was presented in the original orthography from the 19th century, which is outdated today. Here are some examples of orthographical disparities: не ћу 'I will not', now нећу; Црна гора 'Montenegro', now Црна Гора; the appearance of the semivowel њ (see Figure 1), which does not exist anymore in the Serbian alphabet, etc. The editors agreed not to make any corrections in these particular cases.

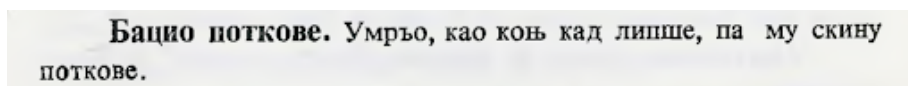


Fig. 1. Semivowel њ.

- Retention of the author’s abbreviations: in the text, there are also some common abbreviations, which are still in use, for instance, н. п., standing for на пример, meaning 'exempli gratia'; т. ј. standing for то јест, meaning 'id est', but in some cases, the orthography has changed, such as in и т. д., now итд. standing for и тако даље, meaning 'et cetera';

- Preservation of the author's self-censorship: in the text, there is an intentional omission of the letters in the middle of obscene words marked with the ellipsis, where each dot represents one omitted letter (see Figures 2 and 3 below). This was an editorial decision of the original author due to puristic tendencies of the public and critique of that time.

**Блажене су многе ручице, ал' су проклете многе  
Г....е. Многи много ураде, али много и поједу.**

Fig. 2. Censored word гузице.

**Ђе се п..и, не да се клањати.**

Fig. 3. Censored word прди.

The first concern of the editors of this digital edition was the preservation of the authenticity of the text, therefore in the succeeding stages of the project we intend to include now missing information about the new orthography, the meaning of abbreviations, and the full form of censored words, while retaining the original text.

There have also been minor interventions, which were similarly motivated by preserving the original characteristics of the printed edition. For example, some parts of text are in verse (see Figure 4), and in order to preserve this information, the editors of the digital edition decided to add a vertical line with spaces before and after it at the end of each verse (Figure 5).

**Вук не вије што је меса гладан,  
Него вије да дружину свије.**

Fig. 4. A rhyme.

```
<entryFree xml:id="630">
  Вук не вије што је меса гладан, | Него вије да дружину свије.
</entryFree>
```

Fig. 5. A rhyme with a vertical line.

## 2.2 Data modeling

The collection of proverbs is organized by the same principle as a dictionary: the items are arranged by the alphabetic order (in this case by the Serbian Cyrillic alphabet, since the text was written in Cyrillic script). The only difference is that, in the case of proverbs, the items are not simple lexical units consisting of one word only, as usual, but multi-word (complex) lexical units with multi-word lemmas. It can be seen in the paper edition that the form of the proverb is

highlighted by the different typographic style, as it is the lemma in the dictionary. Commonly, every new entry is separated by a new paragraph.

The text in the bold style is the form of the proverb – the multi-word lemma, separated from the rest of the text with a full stop. The rest of the text is a note which contains different information. The majority of those notes contain the explanation of the meaning or the usage of a proverb, but, occasionally, they include information about the provenance of the proverb. Most frequently, the information about provenance constitutes a separate sentence (Figure 6), but there are plenty of cases where it is included in the text (Figure 7).

**Али си ми прасицу шишао? Кад који кога зовне: куме,  
а није му кум. У Боци.**

Fig. 6. Information about provenance is separated from the note.

**Ако није, ја цркла, а кад цркнем није ми од преше  
то ђавоље. Заклињу се жене у Црној гори кад што доказују.**

Fig. 7. Information about provenance is included in the note.

The references to other entries in the same collection are also included in the note, mainly introduced by the formula 'Гледај:', meaning 'See:' (Figure 8).

**Забринуо се као курјак у јами. Гледај: Стиди се као  
курјак у рупи.**

Fig. 8. Reference introduced by the formula 'Гледај:'.

Since all this information is optional, the minimal structure of an entry consists only of the multi-word lemma (Figure 9).

**Беговац је беговац, ако не ће имати ни новац; а  
магарац је магарац, ако ће имати и златан покровац.**

Fig. 9. One of the longest proverbs.

The majority of the proverbs consist of only lemma without the note (Figure 10). Diversely, in Figure 11 we can see the maximal structure of an entry, containing all elements of the schema, and in Figure 12 there is a representation of the same entry in XML.

Following the original structure, we have used the TEI P5 standard for dictionaries<sup>3</sup> to annotate data from the collection of proverbs. We decided not to annotate non-linguistic features (levels) of the text, such as division of the proverbs by starting letters, or the end of the page. For linguistic annotation, we used the following labels. We have

<sup>3</sup>See: <https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>

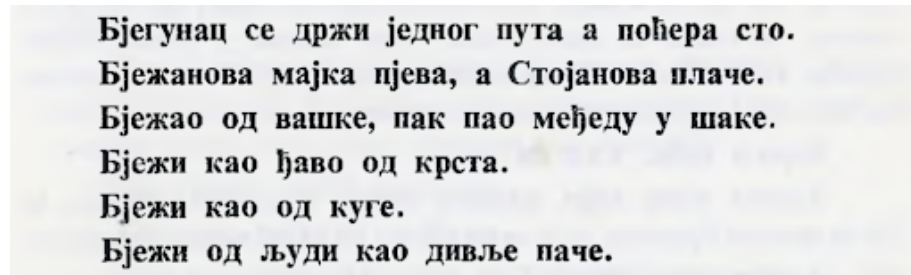


Fig. 10. A series of proverbs without additional information.

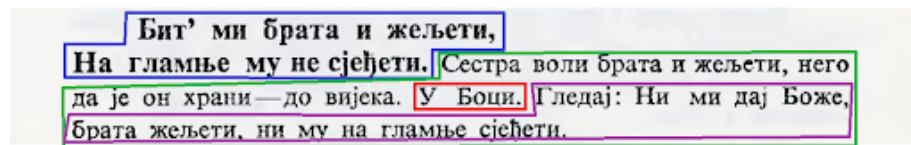


Fig. 11. Example of a fully segmented entry.

chosen the EntryFree element to be the main unit of the macrostructure of the text, containing all the information related to one proverb. Speaking in hierarchical terms, the microstructure of the entry has two top-level constituents: the text of a proverb, contained by the element Form, and the text of a note, contained by the element Note. We have chosen this label because it broadly covers the content and can contain other lower-level constituents we needed: PlaceName and Cross-reference. We decided not to annotate the meaning of the proverb with a separate label such as Sense, because the remaining text in the notes is not always an explanation of meaning, but sometimes an explanation of the usage of the proverb (which would have to be annotated with a different label Usage), or even an anecdote. Distinguishing these cases could not be done differently than by a human decision, which would require additional time and resources, therefore we decided to skip this step for the time being.<sup>4</sup>

```
<entryFree xml:id="8">
  <form type="proverb">Бит' ми брата и жељети, | На гламње му не сјеђети.</form>
  <note>Сестра воли брата и жељети, него да је он храни – до вијека.
    <placeName type='location'>У Боци.</placeName>
    <xr><lbl>Гледај:</lbl><ref><rs type="proverb">Ни ми дај Боже брата жељети, ни му на гламње сјеђети.</rs></ref></xr>
  </note>
</entryFree>
```

Fig. 12. Example of a fully annotated entry.

The first step we took in the annotation was the automatic parsing of the macrostructure. This entailed separating the individual proverbs, marking them with EntryFree labels, and giving each of them a unique ID. This was done using

<sup>4</sup>Compare also the Excerpt of the Encoded Text of the Collection of Proverbs, from the already mentioned electronic edition (see Introduction), given by the author team on the link <https://alas.matf.bg.ac.rs/~cvetana/proverb/append/e-tei.html>

a python script, and the separation was made using a set of heuristic rules that covered most cases. We considered one proverb to start with a capital letter and end with a punctuation mark followed by a newline character.

After this step, we had the complete text of the collection structured with the EntryFree label, but with some errors - still quite a few examples remained to be corrected. There were mistakes in some entries consisting of more than one sentence. In certain cases, the explanation of the proverb was written in a new line and, since it started with a capital letter, the proverb and its explanation would end up separated. Those cases were incorrectly recognized as two separate entries (Figures 13 and 14).

```
<entryFree xml:id="138">
  Што виде четири ока, виђеће и двадесет и четири.
</entryFree>
<entryFree xml:id="139">
  Што знаду двојица, то већ није тајна.
</entryFree>
```

Fig. 13. Example of XML of a wrongly separated proverb from its explanation.

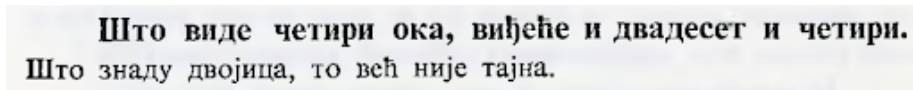


Fig. 14. The same entry in the printed edition.

Another typical example of an error is an entry consisting of two very similar proverbs, which were listed one by one, and connected with the linker Или:, meaning 'Or:'. Those two separate entries were often automatically recognized as one (Figures 15 and 16). In this other case it was not really a mistake of the algorithm, but an oversight of the author of the collection, for he inconsistently used the linkers followed by the colon sometimes as shown, connecting two entries, but prevalently connecting two units inside one entry (as it is in the cross-references above). The correction of these cases was done manually by the annotators.

```
<entryFree xml:id="124">
  Шта је тражио, мало је нашао. Или: Шта је тражио то је и нашао.
</entryFree>
```

Fig. 15. Example of XML of two wrongly connected proverbs.

Annotation was done by high school students, participants of the Tršić summer school. The students had no previous experience, neither with digital humanities nor annotation of XML files, but they picked up the task quite quickly and they had no problems using the Oxygen XML editor.<sup>5</sup> They were given XML with proverbs marked as entryFree elements, and they first had to correct the issues from OCR (see chapter 2.1 above) and the proverb separation.

<sup>5</sup><https://www.oxygenxml.com/>

**Шта је тражио, мало је нашао. Или:  
Шта је тражио то је и нашао.**

Fig. 16. The same entry in the printed edition.

After the proverbs were all correctly separated and listed as EntryFree elements, the next task for annotators consisted of several steps:

- annotating locations with PlaceName label;
- adding location attributes: mention or location;
- annotating notes with Note label.

It was a priority to annotate these entry constituents first since we assumed that other constituents could be annotated automatically without a high mistake rate once these labels were fixed.

Once the annotations were complete, the XML files were shuffled for cross-validation. The annotators were given work from their peers to fix any remaining mistakes and proofread the newly added annotations. This had highly reduced the number of errors in the data. It is, of course, recommendable to have the annotations checked one more time by an expert.

Even after double-checking, we discovered a few mistakes in annotation, two of them were false positive locations. Word челебија, an obsolete Turcism meaning 'young gentleman', was recognized as a location name, probably due to its resemblance to a real location name Келебија. The other case includes a rare word трипуњица meaning 'the day of Saint Tryphon', found in a compound Которска трипуњица, which contains a place name Kotor (Котор). However, this extended named entity does not denote a place name but a holiday name, meaning 'the day of Saint Tryphon of Kotor', which is a spot in time, not in space, and consequently should not be annotated as a place name. This shows us that the location identification task can be challenging even for humans.<sup>6</sup>

### 2.3 Data storage and enrichment

In the initial phase of digitization, our focus was on correctly extracting the proverbs from the OCR text, processing them, and storing them in a database. Processing the proverbs entailed separating the proverb from its explanation and annotating locations. In terms of data enrichment, the information added by the annotators was regarding the type of location, more precisely by adding location type which denotes whether a place name is a mention, or a source, i. e. location where a certain proverb was recorded. This information is stored in the first database, which consists of:

- unique proverb id
- proverb text
- explanation
- locations (type irrelevant)

After the initial processing, done in the annotation phase, was complete, we have added a new layer of information to our proverb location database. This was done by determining the geo-coordinates of the toponyms, both mentioned and cited as a source. For this task, we have initially used a database of geolocations which was previously compiled for

<sup>6</sup>For a detailed discussion on automatic location identification and its challenges see also [Kyriacopoulou et al. 2020]



the dictionary platform Raskovnik.<sup>7</sup> As expected, not all toponyms from the proverbs were present in this database, so we needed to determine the coordinates for the remaining locations.

In order to connect the locations from the proverbs to their respective coordinates, we needed to create a mapping between the locations as recorded in the text and their normalized versions. One aspect of this problem was to lemmatize the inflected location names, while the other involved disambiguating ambiguous locations.

We have first considered creating a lemmatizer for the inflected forms of toponyms, but we soon realized it would be less complicated to manually create a lexicon which contains all the inflected forms and to connect them to their normalized versions, as they can be found in the database with coordinates or the Google API.

Another issue we encountered in this phase was the identification of obsolete forms of toponyms (for instance, Lipisca is the old name for Leipzig). In some cases, the author used elliptical names of location (Novi for Herceg Novi or Karlovci for Sremski Karlovci). We can assume in the time when the proverbs were written, it was implicitly clear what locations were referred to, but we needed to disambiguate them correctly in order to make sure the appropriate coordinates were shown on the map. Additionally, there were some locations which have completely changed, or even disappeared over the course of almost two centuries (for example “The Turkish empire”).

After using our lexicon to disambiguate and normalize the toponyms, it was straightforward to connect them to their respective coordinates. We have stored this information in our second database. The locations database consists of:

- location text (verbatim)
- location name (normalized)
- type
- coordinates
- id of the proverb where toponym is found

## 2.4 Data presentation

The final result of our work is stored in two SQL databases. Out of 5684 annotated proverbs, only 587 contain toponyms. In total, there are 708 annotated instances of locations: 76 mentions and 632 sources cited. When normalized, we are left with 99 specific locations, out of which 50 are sources and 49 are mentions. The most common location cited as a source is Montenegro, as we can see in the table below.

Table 1. Five most common locations

Location	Number of instances
Montenegro	219
Vojvodina	67
Risan	63
Dubrovnik	58
Boka Kotorska	44

Table 2. Toponyms as sources

Location	Number of instances
Kosovo	6
Budapest	5
Bačka	5
Morača	4
Rome	4

Table 3. Toponyms mentioned

<sup>7</sup>See: <http://raskovnik.org/>

In order to present our data in an explicit and comprehensible way, we made a Django web application which shows our data tabularly and allows search and sort. In this way, users can easily filter proverbs and look for particular locations.

In addition, we have also used the Google Maps API to create a visual representation of the locations on a map that differentiate place names only mentioned in proverbs (blue) from ones designating a location where a certain proverb was recorded (red) (see Figure 17).

When a marked location is clicked, you can see the normalized name of the location. For future work, we intend to make the map more interactive and make it possible to get a list of proverbs collected in one location with a click.



Fig. 17. Map of locations from the proverbs.

### 3 CONCLUSION

In conclusion, while the digitization process of Serbian folk proverbs is still ongoing, we believe the approach we have chosen has allowed us to continuously work on having a fully digitized and user-friendly edition once the full annotation of the entire collection is completed. By providing new access paths for search and navigation, we are increasing the use-value of the collection, especially in regards to search abilities and accessibility.

Since the digitization process is still in progress, there are still further possibilities to supplement the digital edition, such as adding new labels in terms of data modeling, like ethnonyms and references, as well as in terms of data enrichment, such as semantic field, marking of omitted and censored proverbs or those that occur with other authors. In

accordance with open science requirements, we plan to make this electronic edition of Serbian folk proverbs available to the public as soon as possible.

We believe that the main benefit of having a digitized edition of Serbian folk proverbs is the possibility to obtain a multi-level insight into the collection, which is not possible with printed editions only. In addition, a digitized edition will give us a broader perspective in terms of any future research of proverbs, and it can serve as a potential model for similar digitization projects in the future. By doing so, the digitized edition will also easily integrate and be compatible with other digitized resources and/or multi-dictionary portals, which would enable its better visibility to scientific researchers and the broader public.

## REFERENCES

- Jovan Deretić. 2004. *Istorija srpske književnosti*. Prosveta, Beograd.
- Cvetana Krstev, Duško Vitas, and Vesna Šatev. 2006. TEI Encoding of Serbian Proverbs.. In *Computer Applications in Slavic Studies, Proceedings of Azbuka.Net International Conference and Workshop* (Sofia, Bulgaria). Institute of Literature Bulgarian Academy of Sciences, 101–114.
- Tita Kyriacopoulou, Claude Martineau, and Markarit Vartampetian. 2020. Extraction and annotation of ‘location names’. *INFOtheca: Journal of Information and Library Science* 19, 2 (2020), 7–25.
- Miroslav Pantić. 1965. *Vuk Stefanović Karadžić i naše narodne poslovice*. Prosveta, Beograd.
- Vuk Stefanović Karadžić. 1965. *Srpske narodne poslovice i druge različne kao one u običaj uzete riječi*. Prosveta, Beograd.
- Toma Tasovac and Snežana Petrović. 2015. Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age* (2015), 384–396.
- Drago Ćupić. 1981. Osvrt na Srpske narodne poslovice sa stanovišta Vukovog kodifikovanja jezičke norme. *Naučni sastanak slavista u Vukove dane* 10 (1981), 61–68.