

# Developing Morpho-SLaWS: An API for the Morphosyntactic Annotation of the Serbian Language

Toma Tasovac<sup>(✉)</sup>, Saša Rudan, and Siniša Rudan

Belgrade Center for Digital Humanities, Belgrade, Serbia  
{ttasovac, sasha.rudan, sinisa.rudan}@humanistika.org

**Abstract.** Serbian Lexical Web Service (SLaWS) is a resource-oriented web service designed to offer multiple functionalities—including morphosyntactic, lexicographic, and canonical text services—to create the backbone of a digital humanities infrastructure for the Serbian language. In this paper, we describe a key component of this service called Morpho-SLaWS, the atomic morphosyntactic component of the service infrastructure. The goal of Morpho-SLaWS is to offer a reliable, programmatic way of extracting morphosyntactic information about word forms using a revised version of the MULTEXT-East specification. As a service-oriented lexical tool, Morpho-SLaWS can be deployed in a variety of contexts and combined with other linguistic and DH tools.

**Keywords:** API design · Service architecture · Morphological lexicon · Serbian language · Digital humanities

## 1 Intro: A Language in Search of an Infrastructure

A recent white paper evaluating the state of the Serbian language technologies has shown that Serbian rates poorly in most categories, including the quantity and availability of lexical resources (Vitas et al. 2012).<sup>1</sup> The lack of easily accessible, open-sourced language resources and ready-made frameworks for quantitative textual analysis that take into consideration the specificities of Serbian morphology and syntax has been a major stumbling block in the development of Serbian Digital Humanities.

Serbian Lexical Web Service (SLaWS) is a resource-oriented web service designed to offer multiple functionalities, including morphosyntactic, lexicographic, and canonical text services that are the backbone of a digital humanities infrastructure for the Serbian language. In this paper, we focus on one key aspect of SLaWS: Morpho-SLaWS, the atomic morphosyntactic component of the service infrastructure and the API for the query-driven extraction of Serbian morphosyntactic data. The web service and the API follow standard practices in the field of web-based language resources, while also making provisions for the peculiarities of contemporary Serbian,

---

<sup>1</sup> At the same time, it is important to keep in mind that Serbian is not the only language suffering from the predicament of underdeveloped resources. “It is estimated that for most European languages, “even the basic resources are not yet available.” (Váradi et al. 2008).

such as active Cyrillic/Latin digraphia and substandard Latin orthographic practices in computer-mediated communication.

Morpho-SLaWS is a flexible and easily pluggable Digital Humanities (DH) tool, developed in conjunction with recent developments in language service infrastructures (Ishida 2006; Váradi et al. 2008; Murakami et al. 2010) and the goals of interoperability, collaborative creation, management, and sharing of language resources by means of distributed architectures (Calzolari 2008). For an under-resourced language like Serbian, Morpho-SLaWS (and the SLaWS Framework in general) represents a major departure in the way language resources are conceptualized, developed, and disseminated. We see Morpho-SLaWS as part of a long-term effort to build an infrastructure, which will encourage programmatic accessibility and manipulability of Serbian textual data in various DH contexts, including text annotation, indexing, cross-referencing, text analysis, and visualization.

## 2 The Infrastructural Turn in Digital Humanities

DH encompass a wide range of scholarly activities ranging from digital philology and creation of digital editions to text mining, distant reading, algorithmic criticism (see Ramsay 2011; Berry 2012; Gold 2012; Liu 2012). As a community of practice, digital humanists deal with electronic text not as a static artifact, but rather as a complex, multi-dimensional and multi-layered datasets that need to be analyzed, annotated, and manipulated in order to produce new knowledge. It should come as no surprise that one of the most important challenges facing Digital Humanities today is how to consolidate and repurpose available tools; how to create reusable but flexible workflows; and, ultimately, how to integrate and disseminate knowledge, instead of merely capturing it and encapsulating it. This technical and intellectual shift is what makes the ‘infrastructural turn’ in Digital Humanities.

The increasing appeal of web services in the context of this infrastructural turn is both technical and social. On the technical level, web services let heterogeneous agents dynamically access and reuse the same sets of data using application programming interfaces (API) and standardized workflows. On the social level, web services help overcome the problem of “shy data,” i.e., data you can “meet in public places but you can’t take home with you.” (Cooper 2010) Designing DH projects in line with the principles of the service-oriented architecture (SOA) is, therefore, an important step in the creation of open scholarly ecosystems.

A growing number of large-scale, international projects is delineating the contours of the infrastructural turn in digital humanities and related fields, in which web services, APIs, and Open Linked Data play a significant role. Large European consortia such as CLARIN and DARIAH coordinate and direct efforts in the realm of digital research infrastructures for language technologies and digital arts and humanities. Initiatives such as the Open Annotation Collaboration (OAC) and the Web Annotation Working Group are working on data models and interoperability specifications for a distributed Web annotation architecture (Hunter et al. 2010; Haslhofer et al. 2011;

Sanderson et al. 2013). While emerging protocols such as the Canonical Text Services (CTS) identify and retrieve XML-structured textual passages of Classical authors using an URN scheme (Smith 2009; Tiepmar et al. 2014).

The trend of open, shareable, and easily accessible “infrastructuralized” data is catching on in individual DH projects as well: Open Siddur’s API (Nosek 2013) retrieves, for instance, Jewish liturgical resources, while Folger Digital Texts offers a simple API to identify and retrieve words, lines, or other segments of Shakespeare plays, concordances as well as the so-called witness scripts for individual characters, i.e., portions of the play that characters witness by virtue of being on stage.<sup>2</sup> The *correspSearch* API provides access to metadata of diverse scholarly letter editions with regard to senders, addresses, as well as places and time of origin.<sup>3</sup> APIs are nowadays used not only to deliver content but also to document and make easily accessible the encoding choices made in creating digital editions (Holmes 2014).

### 3 Morphosyntactic Annotation in the Service of Serbian DH

We know from experience that serious textual work in Digital Humanities, especially with highly inflected languages such as Serbian, cannot be imagined without morphosyntactic analysis. Lemmatization, POS-tagging, and removal of function or most frequent words are basic pre-requirements for a variety of DH practices, including, for instance, annotating a scholarly digital edition, performing a quantitative analysis of a collection of electronic texts, or topic modeling.

In view of both the state of the Serbian language technologies and the infrastructural turn in Digital Humanities, we judged it essential to invest both time and effort in developing a web service that would help with some of those tasks.

As a web service compliant with the REpresentational state transfer (REST) architecture (Richardson and Ruby 2007), Morpho-SLaWS provides a framework for query-based extraction of morphosyntactic data over the network. It follows the principles of Resource Oriented Architecture (ROA): it makes the components of the underlying lexical dataset addressable through URIs; it uses the HTTP GET method to retrieve a representation of the resource; and every HTTP request happens in stateless isolation, which makes backend implementation and architectural integration much simpler.

Compared to conventional XML transport mechanisms such as SOAP, RESTful protocol provides a lighter solution, more suitable for an online infrastructure, including light mobile solutions (see Muehlen et al. 2005). These aspects happen to be particularly important for under-resourced languages, where communities of developers and crowdsourced editors can lead to significant improvements in resource availability and quality.

---

<sup>2</sup> <http://www.folgerdigitaltexts.org/api>.

<sup>3</sup> <http://correspsearch.bbaw.de/index.xql?id=api>.

## 4 Morpho-SLaWS

### 4.1 The Morpho-SLaWS Lexicon

The morphosyntactic dataset that forms the backbone of Morpho-SLaWS was originally developed as part of *Transpoetika*, a bilingualized, WordNet-based Serbian-English dictionary (Tasovac 2009; Tasovac 2012). It has been extended over time and used internally in various contexts including encoding and indexing of literary texts (Tasovac and Ermolaev 2011a; Тасовац and Јермолаев 2012) as well as dictionary backends (Tasovac and Ermolaev 2011b; Чемериќић 2013).

The Morpho-SLaWS Lexicon (MSL) links individual word forms with their corresponding lemmas and morphosyntactic annotation. The MSL tagset relies on the revision of the MULTeXТ-East morphosyntactic specification (Erjavec 2010; Tasovac and Petrović, [forthcoming](#)). The specification defines a formal set of feature structures for annotating salient word-level grammatical properties for each of the languages concerned. The specification also provides a mapping between its feature structures and the so-called morphosyntactic descriptions (MSD)—a compact annotation scheme which can be used in a variety of natural language processing tasks. For instance, the MSD `Ncnpг--n` corresponds to the feature structure consisting of attribute-value pairs `Category = Noun`, `Type = common`, `Gender = neuter`, `Number = plural`, `animate = no`.

The full paradigm of the Serbian noun *писмо* (letter) in the Morpho-SLaWS lexicon looks like this:

| Form    | Lemma | MSD      |
|---------|-------|----------|
| писама  | писмо | Ncnpг--n |
| писама  | писмо | Ncnpа--n |
| писама  | писмо | Ncnpн--n |
| писама  | писмо | Ncnpв--n |
| писама  | писмо | Ncnsг--n |
| писмима | писмо | Ncnpд--n |
| писмима | писмо | Ncnpи--n |
| писмима | писмо | Ncnpл--n |
| писмо   | писмо | Ncnsа--n |
| писмо   | писмо | Ncnsн--n |
| писмо   | писмо | Ncnsв--n |
| писмом  | писмо | Ncnsи--n |
| писму   | писмо | Ncnsд--n |
| писму   | писмо | Ncnsл--n |

The MSL is itself under active development. As of this writing, it consists of 3,948,328 morphological entries for 114,932 lemmas. In comparison, the morphological dictionary of the Serbian language in the LADL/DELA format (Krstev 2008) covers a total of around 4.5 million word forms and 130,000 lemmas (Krstev et al. 2011).

## 4.2 The Morpho-SLaWS Backend

The Morpho-SLaWS backend is implemented as a JavaScript service running in the NodeJS runtime environment, supported by the light Express web application framework and MongoDB persistent storage. MongoDB is a non-relational (NoSQL), schema-less, document-oriented, persistent storage solution, which offers multiple benefits over conventional storage solutions (Wei-ping 2011), such as architectural design patterns greater scalability and standardized solutions for easier replication scenarios.

## 4.3 The Morpho-SLaWS API

### Query Parameters

*Method: GET.*

[https://api.slaws.info/v1/forms/wordForm?fields=apikey,limit,offset,filter,paradigm,transliterate,strict,created\\_since,modified\\_since,form,lemma,ana](https://api.slaws.info/v1/forms/wordForm?fields=apikey,limit,offset,filter,paradigm,transliterate,strict,created_since,modified_since,form,lemma,ana)

The following table outlines a list of possible parameters, indicating whether they are required (y) or not (n), briefly describing their functions, default values and data types (Table 1).

**Table 1.** Morpho-SLaWS parameters

| Parameter     | Req. | Function   | Data Type |
|---------------|------|--|-----------|
| apikey        | y    | API key  | String    |
| limit         | n    | The maximum number of results to return. Default: 10.  | Number    |
| offset        | n    | The pagination of results to return. Default: 0.   | Number    |
| filter        | n    | Filters results based on the MSD notation. * returns all forms, N* returns all nouns, Nc* all common nouns, Nps* proper nouns in the singular, etc. Default: *.  | String    |
| paradigm      | n    | By default, the system returns the MSD for the queried word form. If true, the system returns a full inflectional paradigm of the word form, regardless of whether the requested word form is a lemma or not.<br>/forms/руке?paradigm = false =>руке (gen. sg),<br>руке (nom. pl.), руке (acc. pl.)<br>/forms/руке?paradigm = true =>рука, руке, руци,<br>руку, etc.<br>Default: false | Boolean   |
| transliterate | n    | Transliterate the requested word form from Latin to Cyrillic. This does not affect the output. Default: false.   | Boolean   |
| strict        | n    | If transliterate is set to true and strict is set to false, the system will try to match and offer results for loose transliteration:<br>/forms/реци?transliterate = true&strict = true =>реци<br>/forms/реци?transliterate = true&strict = false =>реци,<br>рећи, речи.<br>Default: true  |           |

(Continued)

**Table 1.** (Continued)

| Parameter      | Req. | Function   | Data Type |
|----------------|------|--|-----------|
| created_since  | n    | Either zero or the Unix timestamp. Default: 0. Returns entries created since a given time and date.  | Number    |
| modified_since | n    | Either zero or the Unix timestamp. Default: 0. Returns entries modified since a given time and date. | Number    |
| form           | n    | Include the requested word form in the response. Default: <code>true</code> .                        | Boolean   |
| lemma          | n    | Include the lemma of the requested word form in the response. Default: <code>true</code> .           | Boolean   |
| msd            | n    | Include the morphosyntactic analysis of the form in the response. Default: <code>true</code> .       | Boolean   |

Most of the parameters and functionalities described in the above table are self-explanatory. Two of them, however, deserve special attention, as they address a peculiarity of contemporary Serbian as an actively digraphic language, which can be natively written in both Cyrillic and Latin alphabets (Magner 2001). The API provides two parameters: `transliterate` and `strict` to accommodate this feature. The former instructs the system to transliterate from Latin to Cyrillic:

```
/forms/prisustvo?transliterate=true =>returns results for
присуство
/forms/čašćavali?transliterate=true => returns results for
чашћавали
```

etc. Unlike the transliteration from Cyrillic to Latin, which is unambiguous, the transliteration from Latin to Cyrillic poses some additional difficulties. Cyrillic graphemes љ, њ, and џ correspond to Latin digraphs: `lj`, `nj`, and `dž`. Unicode does provide single characters for the Latin digraphs, but they are hardly ever used in word processing and in web content. While in majority of cases, the Latin digraphs can be safely transliterated to their monographic Cyrillic counterparts, there are exceptions that require a digraphic Cyrillic representation: for instance, `džak` = `џак` (`dž` = `џ`), but `nadživeti` = `надживети` (`dž` = `дж`).

The system deploys what we call a *maximal transliteration approach*: every Latin-script word is internally transformed into all of its theoretically possible Cyrillic representations: `džak` => `[џ|дж]ак` and `nadživeti` becomes `на[џ|дж]ивети`; the system then returns the results for all the forms that it has encountered in the lexicon.

A further difficulty for the processing of web-based Serbian texts is that a large portion of Serbian speakers in computer mediated communication employs non-standard orthographic practices, most notably the diacritic-free versions of Latin graphemes `č`, `ć`, `š`, `đ`, `ž`, and `dž` (Брборић 2000; Ivković 2013). Keeping in mind that diacritics in the Serbian Latin alphabet are markers of distinct phonemes rather than accent marks, the substandard orthographic practices can interfere with morphosyntactic annotation.

As we saw above, Morpho-SLaWS can transliterate standard orthographic conventions by employing a maximal approach to compensate for the potential graphemic ambiguities. In cases of non-standard orthographic practices, the API accepts an additional parameter, `strict`, to indicate whether the system should try or not to resolve the potential substandard graphemic alternatives.

```
/forms/reci?transliterate=true&strict=true => реци
/forms/reci?transliterate=true&strict=false=>реци,
рећи, речи
```

Setting the parameter `strict` to `false` will search for all possible substandard transliterations and return those that have an entry in the lexicon: in the case of *reci*, the client will receive MSDs for *reci* (реци), *reći* (рећи), and *reči* (речи):

```
реци   редак   Ncmpn--n
реци   редак   Ncmpv--n
реци   река     Ncfsd--n
реци   река     Ncfsl--n
реци   рећи     Vmmp2s-an-n---e
речи   реч     Ncfsd--n
речи   реч     Ncfsг--n
речи   реч     Ncfsi--n
речи   реч     Ncfsl--n
речи   реч     Ncfsv--n
рећи   рећи     Vmn----an-n---e
```

The system accounts for multiple orthographic mappings:

1. the “traditional” non-standard Latin: `c => [cćć]`; `dj => [dj|đ]`; `s => [sš]`; `z => [zž]`; `dz => [dz|dž]`;
2. the “Anglicized” non-standard Latin: `ch => [ćć]`; `cj => ć`; `zh => ž`; `sh => š`; `dzh => dž`;
3. the “telegraphic” non-standard Latin: `cc => ć`; `ch =>ć`; `zz => ž`; `ss => š`; `dzz => dž`;

**Output Formats.** Morpho-SLaWS returns lexicon entries in two formats: as JSON objects and as TEI-XML documents.

*JSON notation.*

```
{ "entry" : {
  "form" : "учитеља",
  "lemma" : "учитељ",
  "msd" : [
    "Ncmsg--y",
    "Ncmsa--y",
    "Ncmpг--y"
  ]
}}
```

*TEI Notation.* Entries from the Morpho-SLaWS Lexicon can also be returned as valid TEI documents, consisting of a `teiHeader` (Fig. 1) and lexicon entries encoded as TEI feature structures. The header, which is a required TEI element, provides basic metadata about the service as well as the full request query (Fig. 2).

```

<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Morpho-SLaWS Analysis of word form "учитеља"</title>
      <respStmt>
        <resp>Service Provider:</resp>
        <name>Belgrade Center for Digital Humanities</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <authority>
        <ref target="http://humanistika.org">Belgrade Center for Digital Humanities</ref>
      </authority>
      <availability>
        <licence>
          <p>CC+BY</p>
        </licence>
      </availability>
    </publicationStmt>
    <sourceDesc>
      <p>Born digital</p>
      <p>
        Your query was:
        <lb/>
        http://api.slaws.dev:8080/forms/учитеља
      </p>
    </sourceDesc>
  </fileDesc>
</teiHeader>

```

Fig. 1. TEI header response

```

</teiHeader>
<text>
  <body>
    <fs>
      <f name="form">
        <string>учитеља</string>
      </f>
      <f name="lemma">
        <string>учитељ</string>
      </f>
      <f name="msd">
        <string>Ncmsg--y</string>
        <string>Ncmsa--y</string>
        <string>Ncmpg--y</string>
      </f>
    </fs>
  </body>
</text>
</TEI>

```

Fig. 2. TEI body response



## 5 Conclusion and Future Work

Currently, Morpho-SLaWS provides atomic access to the morphosyntactic lexicon through the read-only GET interface. It is the first-ever such web service for the Serbian language. It has been successfully tested with ongoing projects at the Belgrade Center for Digital Humanities, including the Transpoetika Dictionary (Tasovac 2012), LitTerra<sup>4</sup>, and Bukvik.<sup>5</sup>

Further work on Morpho-SLaWS will continue in two parallel tracks: technically, we will focus on expanding the scope of the service, on the one hand, to cover batch processing of both plain-text and TEI-encoded XML files; and, on the other, to handle creating, updating and deleting resources. At the same, we will pursue the development of API-based applications in the realm of collaborative editing, crowdsourcing and gamification of annotation tasks and morphosyntactic disambiguation.

## References

- Berry, D.M.: Understanding Digital Humanities. Palgrave Macmillan, Houndmills (2012)
- Calzolari, N: Approaches towards a lexical web: the role of interoperability. In: Proceedings of the First International Conference on Global Interoperability for Language Resources, pp. 34–42 (2008)
- Cooper, D.: When nice people won't share: shy data, web APIs, and beyond. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010) (2010)
- Erjavec, T.: MULTEXT-east version 4: multilingual morphosyntactic specifications, lexicons and corpora. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), European Language Resources Association (ELRA) (2010)
- Gold, M.K. (ed.): Debates in the Digital Humanities. University of Minnesota Press, Minneapolis (2012)
- Haslhofer, B., Simon, R., Sanderson, R., Van de Sompel, H.: The open annotation collaboration (OAC) model. In: Cyberpsychology and Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, pp. 5–9 (2011)
- Holmes, M.: CodeSharing: a simple API for disseminating our TEI encoding. In: Jenstad, J. (ed.): The Map of Early Modern London (2014). <http://mapoflondon.uvic.ca/BLOG10.htm>
- Hunter, J., Cole, T., Sanderson, R., Van de Sompel, H.: The open annotation collaboration: a data model to support sharing and interoperability of scholarly annotations. In: Digital Humanities 2010, pp. 175–78 (2010), [espace.library.uq.edu.au](http://espace.library.uq.edu.au)
- Ishida, T.: Language grid: an infrastructure for intercultural collaboration. In: International Symposium on Applications and the Internet, SAINT 2006 (2006)
- Ivković, D.: Pragmatics meets ideology: digraphia and non-standard orthographic practices in serbian online news forums. *J. Lang. Politics* **12**(3), 335–356 (2013)
- Krstev, C.: Processing of Serbian: Automata, Texts and Electronic Dictionaries. Faculty of Philology, Belgrade (2008)

---

<sup>4</sup> <http://litterra.info/>.

<sup>5</sup> <http://bukvik.litterra.info/>.

- Krstev, C., Vitas, D., Obradović, I., Utvić, M.: E-Dictionaries and finite-state automata for the recognition of named entities. In: Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, pp. 48–56. ACL, Stroudsburch (2011)
- Liu, A.: The state of the digital humanities: a report and a critique. *Arts Humanit. High. Educ.* **11** (1–2), 8–41 (2012)
- Magner, T.F.: Digraphia in the territories of the croats and serbs. *Int. J. Sociol. Lang.* **2001**(150), 11–26 (2001)
- Muehlen, M., Nickerson, J.V., Swenson, K.D.: Developing web services choreography standards —the case of REST vs. SOAP. *Decis. Support Syst.* **40**(1), 9–29 (2005)
- Murakami, Y., Lin, D., Tanaka, M., Nakaguchi, T., Ishida, T.: Language service management with the language grid. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), European Language Resources Association (ELRA) (2010)
- Nosek, J.D.: Open access liturgical resources for judaism. *Theological Librarianship: Online J. Am. Theological Library Assoc.* **6**(2), 63–66 (2013)
- Ramsay, S.: *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, Urbana (2011)
- Richardson, L., Ruby, S.: *RESTful Web Services*. O'Reilly Media Inc., Sebastopol (2007)
- Sanderson, R., Ciccarese, P., Van de Sompel, H.: *Open Annotation Data Model*. W3C Community Draft 8 (2013)
- Smith, N.: Citation in Classical Studies. *Digital Humanities Quarterly* (2009). <http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html>
- Tasovac, T.: More or less than a dictionary? wordnet as a model for Serbian L2 lexicography. *Infotheca: J. Inf. Librarianaship* **10**(1–2), 13–22 (2009)
- Tasovac, T.: Potentials and challenges of WordNet-based pedagogical lexicography: the transpoetika dictionary. In: Granger, S. (ed.) *Potentials and Challenges of WordNet-Based Pedagogical Lexicography: The Transpoetika Dictionary*. *Electronic Lexicography*, pp. 237–58. Oxford University Press (2012)
- Tasovac, T., Ermolaev, N.: Encoding diachrony: digital editions of serbian 18th-century texts. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) *TPDL 2011*. LNCS, vol. 6966, pp. 497–500. Springer, Heidelberg (2011a)
- Tasovac, T., Ermolaev, N.: A User-Centered Digital Edition of Vuk Stefanović Karadžić's *Lexicon Serbico-Germanico-Latinum*. *Digital Humanities 2011* (2011b). <http://xtf-prod.stanford.edu/xtf/view?docId=tei/ab-297.xml;query=:brand=default>
- Tasovac, T., Petrović, S.: *MULTEXT-East Revisited: Serbian Morphosyntactic Tags in Action* (forthcoming)
- Tiepmar, J., Teichmann, C., Heyer, G., Berti, M., Crane, G.: A new implementation for canonical text services. In: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). ACL, Stroudsburch (2014)
- Zhu, W.-P., Li M.-X., Huan, C.: Using MongoDB to implement textbook management system instead of MySQL. In: *IEEE 3rd International Conference on Communication Software and Networks (ICCSN)* (2011)
- Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., Stanojević, M.: The serbian language in the digital age. In: Rehm, G., Uszkoreit, H. (eds.) *META-NET White Paper Series*. Springer, Heidelberg (2012)
- Váradi, T., Wittenburg, P., Krauer, S., Wynne, M., Koskenniemi, K.: CLARIN: common language resources and technology infrastructure. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), European Language Resources Association (ELRA) (2008)

- Брборић, В.: О Језичком расколу. Социоллингвистички огледи I. Београд и Нови Сад: ЦПЛ-Прометеј (2000)
- Тасовац, Т., Јермолаев, Н.: Дијахронијски приступ дигиталним издањима српских текстова 18. века. In: Вранеш, А. (ed.) Дигитализација културне и научне баштине 71–88. Филолошки факултет, Београд (2012)
- Чемериќић, Д.: Збирка речи из Призрена ДимитријаЧемериќића. Центар за дигиталне хуманистичке науке, Препис.org. Београд (2013)