Impressum

FOR THE EDITOR:
Prof. Aleksandar Jerkov, PhD office@unilib.bg.ac.rs
University Library "Svetozar Marković"
Faculty of Philology, University of Belgrade

EDITOR:
Faculty of Philology, University of Belgrade
University Library "Svetozar Marković"
Serbian Academic Library Association

EDITORS OF THIS ISSUE:
Prof. Cvetana Krstev, PhD cvetana@matf.bg.ac.rs
Prof. Ranka Stanković, PhD

MANAGING EDITOR:
Aleksandra Trtovac, PhD aleksandra@unilib.rs
University Library "Svetozar Marković"

EDITOR OF ONLINE EDITION:
Jelena Andonovski, PhD andonovski@unilib.rs
University Library "Svetozar Marković"

ISSN 1450-9687 (print edition)
ISSN 2217-9461 (online edition)

Belgrade, Vol. 21, No. 2, December 2021

WEB PORTAL:
Jelena Andonovski, PhD
University Library “Svetozar Marković”

LECTOR FOR ENGLISH:
Prof. Ivan Obradović, PhD
JeRTeH

DESIGN AND PREPRESS:
Branislava Šandrih Todorović, PhD
Faculty of Philology, University of Belgrade
Info*theca* Team

REDACTOR OF REFERENCES AND UDC:
Nataša Dakić
University Library “Svetozar Marković”

DOI REDACTOR:
Miloš Utvić, PhD
Faculty of Philology, University of Belgrade

JOURNAL REDACTION:
Journal Info*theca*
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infothecajdh@gmail.com

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Foreword

The papers selected for this issue of the Infotheca journal present contributions of the Serbian team in the COST action CA16204 *Distant Reading for European Literary History* (https://www.distant-reading.net). The project started in 2017 with a plan to create a network of researchers which would develop resources and methods within the Distant Reading paradigm – usage of computational methods in the analysis of a large group of literary texts. One of the ultimate goals was to establish good, innovative methods and techniques of computational analysis of texts from this specific group.

One of the most important achievements of this action is the development of a multilingual corpus, European Literary Text Collection - ELTeC which contains 100 novels first published in the period 1840-1920 for many European languages, including Serbian. All texts in this corpus were selected, prepared and annotated using the common principles.

The Serbian team constitutes researchers from the University of Belgrade (Faculty of Philology and Faculty of Mining and Geology), University Library "Svetozar Marković" and Society of Language Resources and Technologies JeRTeh. In their work they received an unconditioned support of numerous researchers gathered by this COST action, and specifically of their leader Prof. Dr. Christof Schöch. The results we achieved would not be possible without the help of numerous volunteers from the JeRTeh society. We are grateful to Prof. Ivan Obradović for carefully reading all papers in this issue and suggesting valuable improvement.

We are grateful to Wikimedia Serbia that helped the printing of this issue of Infotheca through the project "WikiELTeC – Wikidata about old Serbian novels from collection ELTeC".

Leaders of the Serbian team and
Co-editors of this Infotheca issue
Prof. Dr. Cvetana Krstev
Prof. Dr. Ranka Stanković

Contents

Scientific papers

Reviews

Scientific paper

The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection

UDC 004.738.5:027
DOI 10.18485/infotheca.2021.21.2.1

ABSTRACT: In this paper we present challenges and solutions in preparing the Serbian part of ELTeC collection, which contains 100 novels written and first published between 1840 and 1920. In the absence of a systematic digital library of Serbian literature this was done from scratch: first, it was necessary to find out which novels existed and could be used, then they had to be retrieved, scanned, corrected and annotated. All this was achieved thanks to enormous efforts of an army of devoted researchers-volunteers. We analyze the results of these efforts and how they fit to the Action's anticipated outcome.

KEYWORDS: Serbian literature, digitization, corpus annotation, ELTeC collection, library catalogue.

PAPER SUBMITTED: 28 November 2021
PAPER ACCEPTED: 06 December 2021

Aleksandra Trtovac
aleksandra@unilib.rs
Vasilije Milnović
milnovic@unilib.rs
University Library
"Svetozar Marković"
Belgrade, Serbia

Cvetana Krstev
cvetana@matf.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

1 Sampling Criteria for ELTeC

The main goal of the COST Action CA16204 *Distant Reading for European Literary History* is to create a large benchmark corpus of literature from the period 1840-1920 for different computational distant reading methods, as well as for corpus annotation and analysis. The creation of such an ambitiously conceived multilingual corpus required careful preparation. Thus, criteria for selecting a work for the corpus were first specified, namely, the first set of sampling criteria, known as eligibility criteria:[1]

[1] The base reference for sampling criteria was the WG1 Sampling Proposals document, which was agreed by WG1 in 2018. More detailed information can be found in the document ELTeC: corpus composition and extension collections, prepared by the Task Force on Corpus Composition Criteria and approved by the Working Group in August 2020.

- Only novels are eligible, i.e. narrative prose (novels, novellas or longer stories) at least 10,000 words long, which means that works such as travelogues, essays, biographies, autobiographies, historical writings and the like are not considered.
- The first edition of the work should be from the period 1840 to 1920, including these two years.
- The work should be originally written in the language of the sub-collection into which it will be included, which means that translations are not taken into account.
- The work should be published in Europe no later than ten years after its first edition. This provision applies primarily to works in, for instance, English or Portuguese that may have been published for the first time in the United States or Brazil.
- Preference is given to works that were published as books in the specified period, and not in sequels in serial publications.

In addition to these mandatory criteria, additional conditions were set for the composition of each sub-collection, which should, on the one hand, provide diversity of represented texts and, on the other hand, enable comparative analysis of sub-collections and application of key methods for their statistical analysis. These additional criteria for desirable corpus balance are as follows (more about these criteria in Section 5):

- *Collection size*: the sub-collection should contain 100 works that qualify as novels (according to the previously mentioned eligibility criteria).
- *Gender of authors*: the sub-collection should contain works written by both male and female authors, and preferably 30%, but at least 10% of the selected works were written by women.
- *Reprint count*: the sub-collection should contain works from the canon, i.e. well known to the general public, as well as completely unknown and forgotten works. It was decided to take the number of repeated editions of a work as a measure of its canonicity, so the first category includes all works that in the period 1970-2010 had at least two additional editions, while all the others belong to the second category. There should be at least 30% of the latter, but not more than 70%.
- *Even coverage of the period 1840-1920*: The selected time period of the first edition of the work is divided into 4 periods lasting 20 years (only the last period covers 21 years). These time periods should be evenly represented in each sub-collection and each should optimally have 20-25 works.

- *Length of works*: Works are divided according to their length into short (with 10,000-50,000 words), medium length (50,001-100,000) and long (with more than 100,000 words). Each sub-collection should contain at least 20% of works of each length, and ideally 30-40%.
- *Number of novels per author*: Each sub-collection should contain 9 to 11 authors, represented with exactly 3 works, while all other works should be written by different authors to ensure diversity. If for some collections have a difficulty in meeting the other balance criteria, a limited number of authors can be represented with two novels.

In the continuation of this paper we will talk about the development of the Serbian ELTeC sub-collection, dubbed SrpELTeC. In Section 2 we will discuss the importance of the period 1840–1920 for the Serbian literature, and especially the importance of SrpELTeC for reconsidering existing canons. We will continue in Section 3 by explaining which methods we have used to populate the list of eligible novels. Next, we will describe the path we took to get from the title of a work to its electronic edition complying to the prescribed rules (Section 4). In Section 5 we will analyse the extent to which SrpELTeC has met the balance criteria. Finally, in Section 6 we will give some concluding remarks and highlight the importance of achieved results.

2 The Significance of the Period 1840–1920 for the Serbian Literature

It has already been noticed in the time of postmodernism, when it comes to Anglo-literature, that traditional literature has been found to have been written by "dead white males" to serve the ideological aims of a conservative and repressive Anglo hegemony [. . .] In an array of reactions against the race, gender, and class biases found to be woven into the tradition of Anglo lit, multicultural writers and political literary theorists have sought to expose, resist, and redress injustices and prejudices (Stevenson 2007).That is why the success of multiculturalist critique followed quite logically: reading lists were broadened to include more works by women, minority writers, peripheral literatures, historical flexibility/contingency of canon (alternative canons coexisting). Literary symbolization and interpretation of basic existentialia, long-lasting mental structures, multifacetedness, presuppositional complexity, semantic coherence, plurirelation of meaning, archetypal structure, aporeticism. . . are certainly reasons for the selection of certain literary works into the canon, but do not say much about the mechanisms of selection, on institutions and roles that ensure the durability of these texts and their adaptation to ever-changing historical circumstances (Juvan 2019).

That is why it is always necessary to search for principles that led to a selection of texts, comments and explanations to a particular reading audience, while the rest of literary production remains in the blind corner (Juvan 2019). One of the promises of digital humanities scholarship, has been the potential, even the necessity, of moving beyond canons. This can be seen very plastically in Serbian literature.

One of the promises of digital humanities scholarship, is the potential, even the necessity, of moving beyond canons. The Serbian literature is a good example for that. The period determined in this project, contrary to languages with a longer literary tradition, coincides with the introduction of the novel as a genre in the Serbian literature after the language reform of Vuk Stefanović Karadžić. Since one of the main activities within the D-Reading COST action is the production of the ELTeC collection, our first step was the selection of novels that meet the eligibility criteria (see sections 1 and 5) and retrieval of their first editions. In Serbian literature, the indicated period also coincides with the epoch of realism. However, the characteristic of Serbian realism was the predominance of long stories over novels. At the same time, the term *novella* in the Serbian realistic tradition – unlike the Anglo-American one – referred to a long story. That is why some long stories were selected in this corpus, especially since they have significantly more words than the required 10,000.

In addition, care was taken to include some lesser-known writers, outside the official literary canon, or certain works that represent a kind of alternative to the dominant flow of Serbian literature. In that sense, in addition to well-known and recognized Serbian writers of this period, such as Jakov Ignjatović, Milovan Glišić, Laza Lazarević, Stevan Sremac or Bora Stanković, lesser-known writers were included in the corpus, some of whom are actually extremely important. An alternative canonization of Serbian literature would certainly count on a writer like Lazar Komarčić – the first Serbian science fiction author,[2] or Dragutin Ilić, who is the writer of the first science fiction drama in the world – *Posle milijon godina* (After a Million Years) (1889) – published six years before the Time Machine novel by H. G. Wells. Also, we insisted on female authors and included some very important examples

[2] SrpELTeC collection contains three novels by Lazar Komarčić, while his science fiction novel *Jedna ugašena zvezda* (An extinguished star) published in 1902 is included in SrpELTeC-ext.

of women's writing in Serbian literature of that period. For example, one of the selected novels is the novel *Nove* (New Women) (SRP19120) by Jelena Dimitrijević. This novel fell into oblivion after the author's death, and only recently have researchers discovered not only the significance of this novel, but also the fact that it is a true masterpiece of Serbian literature.

To illustrate that the mainstream literary critics have so far neglected or underestimated the women authors, we could cite Prof. Jovan R. Deretić, a prominent historian of literature, who in (Деретић 1981) mentions 30 out of 66 different authors represented in SrpELTeC (see Subsection 3.2 in (Krstev 2021) in the same issue). In this same work, Deretić mentions one of 4 female authors represented in SrpELTeC, Isidora Sekulić, just to say that "her novel remained an unsuccessful attempt".[3] In his other work, Jovan Deretić mentions 33 out of 66 different authors in SrpELTeC (Деретић 1983), among them two more female authors, Jelena Dimitrijević and Milica Janković, in one short, not very favorable paragraph: according to the author "The older of them, Jelena Dimitrijević in her short stories, travel letters, and novel "New women" mostly described oriental Muslim world, particularly the life of Turkish women, while Milica Janković wrote subjective, confessional prose with a lot of elements of old-fashioned sentimentality."[4]

3 Populating the List of Serbian Novels 1840-1920

Publications in the field of the history of literature generally contain knowledge about the most important writers and their most significant works. However, in its history, Serbian literature also has writers who have published only one or just a few novels, or important writers with only one edition of some of their less significant works. The COST action D-Reading takes into account such cases making forgotten writers and/or works an equally important part of the ELTeC collection (see Section 5). It was therefore necessary

[3] *... а док је једини роман И. Секулић* Ђакон Богородичине цркве *(1920) остао само неуспео покушај.* (Деретић 1981, 252) The novel in question is *Đakon Bogorodičine crkve* (Deacon of the Church of the Mother of God) (SRP19190).

[4] *Старија од ње Јелена Димитријевић (1862-1945) у својим приповеткама, путничким писмима и роману* Нове *приказивала је највише оријентални, муслимански свет, нарочито живот турских жена, а млађа Милица Јанковић (1881-1939) писала је субјективну, исповедну прозу с доста елемената старинске сентименталности (*Исповести*, 1913).*(Деретић 1983, 489)

to research in detail the catalogs and bibliographies that provide lists of the entire literary opus for the given period.

The Mutual catalog of the Republic of Serbia,[5] on the COBISS+ platform,[6] contains over 3 million bibliographic records created by over 230 libraries. Due to the contribution of member libraries, and primarily to the Matica Srpska Library, whose entire collection is in its electronic catalog, it was possible to find the necessary information in both coded and bibliographic data within catalog records quickly and easily, using various criteria.

However, before starting to search, it was necessary to devise a strategy how to search for relevant records, in order to avoid duplicate hits when submitting new queries. We decided to search using coded data first, and only then using bibliographic data from the bibliographic record format (Bacotić and Ristović 2020). If we take a look at the structure of the bibliographic record in the COMARC/B format, we will notice that the field `105 - textual material, monographic`, in the subfield `f - literature code` contains the code "a" for "fiction". Since we wanted to find novels in Serbian that were not translated from other languages, and were published between 1840 and 1920, we entered the following query in the expert search:[7]

```
lc=a* and (la=(srp or scc or scr) not lo=*) and py=1840:1920
```

With this query we got 396 hits from the Mutual Catalog. In some cases, we wanted to make the search more concrete by adding the specific author:

```
lc=a* and (la=(srp or scc or scr) not lo=*) and py=1840:1920
and au=ranković, svetolik*
```

This query gave us as the result 4 novels written by Svetolik Ranković and first published between 1840 and 1920.

We noticed that in some cases catalogers were not sure whether certain work was a novel or an extensive short story, so we checked also all records that in the field `105`, subfield `f`, contained the code "`f`" for the short prose. It turned out that some of these works were actually novels – not only because of their size, but because they had other necessary features. This was the

[5] Mutual catalog of RS

[6] COBISS+

[7] `lc` stands for literary genre, `la` for language, `lo` for language of the original, `py` for year of publication, `au` for author, `ti` for title. More detas about the command search can be found in COBISS3 - Catalogization - Command Search.

case, for example, with *Očevi i deca* (Fathers and Sons) by Stevan Mamuzić published in 1898.[8]

Since in retrospective cataloging it is common to make an abbreviated and not comprehensive bibliographic description, we had to keep in mind that each record might not contain the code for the literature in the field `105`, subfield `f`. Thus, we enhanced the research of the Mutual Catalog by using the bibliographic data entered in field `200 - title and statement of responsability data`, subfield `e - other title information`. Namely, subtitles of Serbian novels from the period 1840 to 1920 often contain words such as “roman” (novel), “pripovest” (narrative), “povest”, “pripovetka” (story), “novela” (novella) (see also (Krstev 2021) in the same issue). For this reason, subsequent searches looked for this words in the `title` element when forming the query. We have taken into account both old and modern orthography, the Ijekavian and Ekavian pronunciation of these words, as well as the new (`srp`) and old codes (`scc` and `scr`) for the Serbian language. In order to avoid information that we already extracted, we excluded records that contained coded data in the subfield `105$f` – values “`a`” for novels or “`f`” for short prose.

```
(ti=roman* and (la=(srp or scc or scr) not lo=*) and
py=1840:1920) not lc=(a or f)
```

title	Hits
`ti=istoriski roman*`	3
`ti=istorijski roman*`	6
`ti=pripovest*`	1
`ti=pripovijest*`	12
`ti=povest*`	37
`ti=povijest*`	31
`ti=pripovetka*`	212
`ti=pripovijetka*`	4
`ti=pripovedka`	30
`ti=novela*`	65
`ti=priča*`	47

Table 1. Results of queries using `title` subfield

[8] This novel did not enter SrpELTeC; it will be published in SrpELTeC-ext.

This query produced 86 hits. In subsequent queries we just changed values of the subfield `ti` as presented in Table 1.

We sorted the obtained results according to the surname and the name of the author, and removed from the list authors and novels that were already on the list of novels fulfilling the eligibility criteria. For the queries with subtitles containing "pripovetka", "novela", "priča" we excluded hits that definitely could not reach 10,000 words in size (e.g. less than 50 pages). For remaining hits we reviewed publications *de visu* to determine whether they have a clear literary structure of a novel, rejecting non-fiction narratives.

The next problem was to determine the year when a novel from the list was first published. In the 19th and early 20th century, novels by Serbian writers were first published in sequels in literary magazines and newspapers, even in daily newspapers (see (Krstev 2021) in the same issue, Subsection 3.4). The Mutual Catalog of the Republic of Serbia does not provide information on the first editions in periodicals, nor are the novels in the sequels cataloged in this Mutual Catalog. Fortunately, editions of novels in the form of monograph publications usually contain information that they were reprinted from a certain literary magazine or newspaper, or this information is given in the preface. However, the monograph editions usually do not specify the year(s) when the novel was published in sequels. In such cases the year of the publication of a novel in the book form was used as an orientation in finding the relevant year of the first edition in literary periodicals and newspapers. The only way to determine accurately and unambiguously the year of the first publication was to browse through the old periodicals.

Fortunately, since old periodicals are extremely important and valuable, most of them are digitized and are part of digital collections of the University Library "Svetozar Marković",[9] the Matica Srpska Library[10] and the National Library of Serbia.[11] The digital collection of periodicals of the University Library "Svetozar Markovic" enables full text search by keywords (Тртовац, Андоновски, and Дакић 2021), while the collections of the other two libraries cannot be searched in that way. They contain only scanned pages, which the users can flip through in the browser.

In some cases, the novels printed as monographs did not contain information in which magazine or newspapers they were first published in sequels. In such situations, in order to find information on the first publications, it was necessary to study the entire opus of the authors, e.g. whether they

[9] Digitized Historical Newspapers

[10] Matica Srpska Digital Library

[11] Digital National Library of Serbia

were members of editorial boards of some literary magazine, etc. This implied browsing a large number of periodical titles and publishing years, and a research of authors' biographical data; however, the results were fruitful. In only one case we failed to find the full information because no library had the complete years of the magazine in which the novel was published in sequels: *Dve sestre: samoubistvo jedne švalje* (Two Sisters: the Suicide of a Seamstress) by Božo Savić (SRP19031) published as a monograph in 1903, reprinted from the "Mali žurnal".

4 From Title to e-edition

4.1 Digitization

Scanning of works selected for SrpELTeC was performed in libraries with which cooperation was established and which had the required copies. However, most of the works were found and scanned at the University Library "Svetozar Marković" (more about libraries that participated in this project in (Krstev 2021) in this issue, Section 3.6).

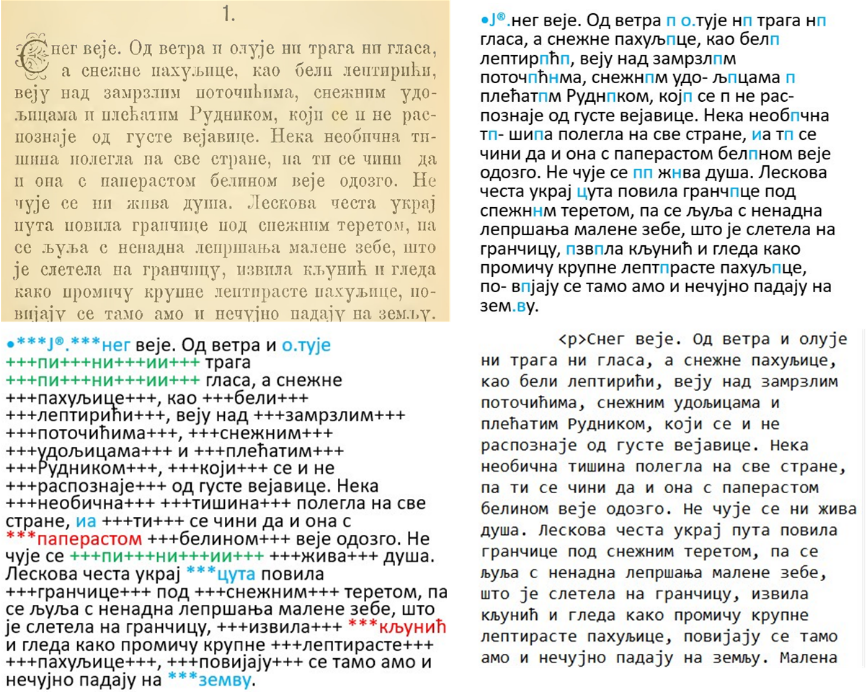
It was the responsibility of the Digitization Department of the University Library to scan the selected material. For these purposes, a Robotic book scanner,[12] type RBS 3.0 from 2014, was used. The output scans were in JPG format with a resolution of 300dpi. From the technical point of view, all settings, such as contrasts, appropriate light, and binarization have been set so that the optimal result is obtained. Optical character recognition was performed with ABBYY FineReader 12. In some cases, due to extremely poor quality of the first edition printing and bad character recognition as a result, a later edition that was also published before 1920 was scanned instead,as is the case with the novel in (Figure 1 (a)). The material prepared in this way was converted into full PDF format. The resulting scans were uploaded to the University Library cloud and shared with the project coordinator.

4.2 The OCR Errors Correction

Scanning was followed by further processing, which was done in several steps:

– The quality of the text obtained by character recognition varied from text to text. In some, rare cases, it was quite good, in some acceptable, and in some, also rare cases, so bad that it could not be used (Figure 1 (b)).

[12] Robotic book scanner



1.

Снег веје. Од ветра и олује ни трага ни гласа, а снежне пахуљице, као бели лептирићи, веју над замрзлим поточићима, снежним удољицама и плећатим Рудником, који се и не распознаје од густе вејавице. Нека необична тишина полегла на све стране, па ти се чини да и она с паперастом белином веје одозго. Не чује се ни жива душа. Лескова честа украј пута повила гранчице под снежним теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, извила кљунић и гледа како промичу крупне лептирасте пахуљице, повијају се тамо амо и нечујно падају на земљу.

•J®.нег веје. Од ветра п о.тује нп трага нп гласа, а снежне пахуљпце, као белп лептирпћп, веју над замрзлпм поточпћнма, снежнпм удо- љпцама п плећатпм Руднпком, којп се п не рас- познаје од густе вејавице. Нека необпчна тп- шипа полегла на све стране, иа тп се чини да и она с паперастом белпном веје одозго. Не чује се пп жнва душа. Лескова честа украј цута повила гранчпце под спежнмм теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, пзвпла кљунић и гледа како промичу крупне лептпрасте пахуљпце, по- впјају се тамо амо и нечујно падају на зем.ву.

•***J®.***нег веје. Од ветра и о.тује +++пи+++ни+++ии+++ трага +++пи+++ни+++ии+++ гласа, а снежне +++пахуљице+++, као +++бели+++ +++лептирићи+++, веју над +++замрзлим+++ +++поточићима+++, +++снежним+++ +++удољицама+++ и +++плећатим+++ +++Рудником+++, +++који+++ се и не +++распознаје+++ од густе вејавице. Нека +++необична+++ +++тишина+++ полегла на све стране, иа +++ти+++ се чини да и она с ***паперастом +++белином+++ веје одозго. Не чује се +++пи+++ни+++ии+++ +++жива+++ душа. Лескова честа украј ***цута повила +++гранчице+++ под +++снежним+++ теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, +++извила+++ ***кљунић и гледа како промичу крупне +++лептирасте+++ +++пахуљице+++, +++повијају+++ се тамо амо и нечујно падају на ***земву.

```
          <p>Снег веје. Од ветра и олује
ни трага ни гласа, а снежне пахуљице,
као бели лептирићи, веју над замрзлим
поточићима, снежним удољицама и
плећатим Рудником, који се и не
распознаје од густе вејавице. Нека
необична тишина полегла на све стране,
па ти се чини да и она с паперастом
белином веје одозго. Не чује се ни жива
душа. Лескова честа украј пута повила
гранчице под снежним теретом, па се
љуља с ненадна лепршања малене зебе,
што је слетела на гранчицу, извила
кљунић и гледа како промичу крупне
лептирасте пахуљице, повијају се тамо
амо и нечујно падају на земљу. Малена
```

Figure 1. Processing steps (example of the first page of the novel *Hadži Đera* by Dragutin Ilić (SRP19040): (a) scanned page (top left); (b) OCR (top right) – in blue are misread characters; (c) automatic correction (bottom left) - in blue are words that could not be corrected, in red words missing from the e-dictionaries, in green words for which there are multiple candidates for correction, while words enclosed with '+++' are corrected words; (d) corrected text after reading (bottom right).

All texts that were not rejected had to be corrected in any case, which was first done automatically, as described in (Krstev and Stanković 2020, 63–68). In short, the automatic correction system looks for each word from the text in the electronic morphological dictionary of the Serbian language (Krstev 2008), and presuming that all words not found in the dictionary are incorrect, replaces them with a one or more words from the same dictionary that could produce the incorrect word as a result of some common recognition error. These common errors are not always the same, so the system had to be adapted to each specific text. One of the most common errors in reading characters in the Cyrillic text is to replace "и" (i) with "п" (p) (and vice versa), "п" (p) with "н" (n) (and vice versa) and "н" (n) with "и" (i) (and vice versa). The set of potential substitutions of an incorrect word can be empty (meaning that the word may have been read correctly, but is not recorded in the dictionary), or contain one or more candidates (Figure 1 (c)).

- In a large number of cases, the scanned text became readable only after the automatic correction described in the previous point. Each text corrected in this way was then read by a reader-volunteer, who compared the text with the original, corrected the remaining errors and chose the right candidate where more were offered (more about volunteers readers in (Krstev 2021) in this issue, Subsection 3.6). The instruction to the readers was that the text should remain true to the original, that is, that they should not correct errors from the printed version, and especially not make adaptations to modern orthography (e.g. some words earlier written separately are now joined into one word, lowercase/uppercase letters are differently used, etc.).
- The third and last control consisted of comparing the text with the electronic dictionary of Serbian again; unrecognized words represented either errors that were missed in the previous step, in which case they were corrected, or words that were missing from the electronic dictionary, in which case the dictionary was enriched with them, or some specificity of the text in terms of spelling or vocabulary, which were left unchanged (Figure 1 (d)).

4.3 Text Annotation

The work on the ELTeC corpus envisages a basic annotation for all subcollections, the so-called level-1 annotation. The annotation consists of marking the basic structural elements of the text (chapters and other units) and some basic textual elements. The annotation was done in accordance with the TEI recommendations (TEI Consortium 2021), where, from a rich set of elements that define these recommendations, only a small subset was selected as mandatory or allowed.

The structural elements are the following:

- The basic element is `<div type = "chapter">`, which is inserted at the beginning of each chapter. If the novel is divided into parts, then each of them is marked with `<div type = "group">`. Chapters and parts can have one or more headings for which `<head>` is used. At the end of a chapter or the whole novel additional information can be tagged with `trailer` – in SrpELTeC it was mostly used for the date of writing, when provided by the author.
- The elements `<front>` and `<back>` are used for front and back matter, respectively. In SrpELTeC `<front>` was used for title pages for all scanned novels (unless they were missing), while `<back>` was used mostly for notes, that is, for footnotes that appeared in a text. These notes were linked with the reference point in the text via the `<ref/>` element. The notes are envisaged for authorial footnotes. In almost all novels in which footnotes appear it was clear that they were authorial. For the remaining few, it was not clear who wrote them – authors or editors/publishers, so they were annotated as notes as well.
- If something resembling a poem appears in the novel – in the form of separate groups of lines – the tags `<quote>` for the whole "poem" and `<l>` for individual lines are used. The `<quote>` element can also be used for other citations (eg. to cite parts of another text, an epigraph at the beginning of the whole text or a chapter).
- To mark the beginning of a new page in the printed work, a special tag is used, e.g. `<pb n = "55" />`. These tags are very useful for correcting text, as well as for parallel display of scanned and read text, as is the case in the digital library of the University Library "Svetozar Marković" (for more about this platform see (Stanković, Škorić, and Popović 2021) in this issue, Section 2).
- The chapter subdivisions are separated by the `<milestone/>` tag; in the text, such subdivisions are indicated by lines, one or more asterisks or a vignette.
- In order to indicate omitted material the tag `<gap/>` is used. When preparing texts for SrpELTeC, it was mostly used to indicate a position at which there was an illustration in the original text, e.g. `<gap unit="graphic"/>`.

- Dividing text into paragraphs marked with `<p>` tags is mandatory, and for SrpELTeC they were added automatically, based on the hard end of the line that the optical character reading generally retains, and readers checked during text correction.

The following text elements are allowed:

- If a title is mentioned in a text (usually given in italics or enclosed by quotation marks) – the title of a newspaper, book, theater play, etc. – it was marked with the tag `<title>` by a reader.
- If a passage in a foreign language appears in the text (it can also be in italics, but not necessarily), it is marked with the `<foreign>` tag to which the language attribute must be added, e.g. `<foreign xml:lang="FR">`. In these cases readers had to retype the text in foreign language, because OCR, which was set to work with only Cyrillic script, could not recognize it.
- A text segment that is somehow highlighted (in italics, bold, underlined, larger font, character spacing, etc.), and does not belong to any of the previous cases, is marked by the reader with the label <hi> (highlighted).

Numeric data about the use of these tags in SrpELTeC is given in (Stanković et al. 2021) in this issue. A metadata header is required for each text annotated in accordance with TEI recommendations, as is the case with all ELTeC texts. It was agreed which header elements are mandatory for all ELTeC texts, so that the headers of all collections are uniform. More about TEI headers for ELTeC, especially for SrpELTeC can be read in (Krstev 2021) in this issue.

5 Compliance of the Serbian Collection with the Corpus Composition Criteria

In order to enable assessment of the degree of compliance of a language subcollection with the composition criteria, a measure is constructed that takes into account all criteria and their relative importance.[13] Its calculation will be presented in the following paragraphs, and illustrated by data from the SrpELTeC.

Collection size factor f_{cs}, where N_{novels} is the number of novels in the collection, and it has the maximum value 10 when the collection has 100 novels, which is the case for Serbian (Figure 2, left).

[13] Details about E5C measure can be found in the E5C Discussion Paper.

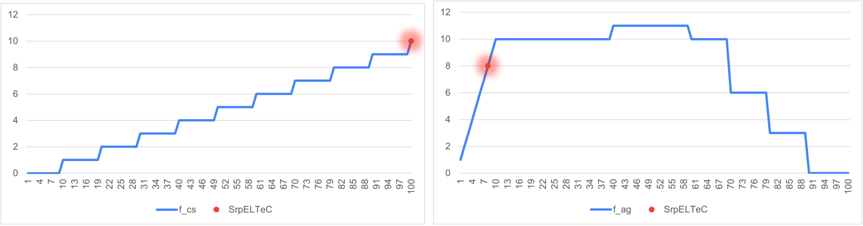


Figure 2. Collection size factor $f_{cs} = f(N_{novels})$ (left); Author gender factor $f_{ag} = f(AG)$ (right). Red dots are factor values for SrpELTeC.

Author gender factor f_{ag}, where AG is the percentage of novels written by female authors. There should be at 10% of such novels, ideally 40-69%, but not more than 70%. This factor has the maximum value 11 when $AG \in [40, 70)$, while its value for SrpELTeC is 8, because there are 8 novels written by women in the 100 novels collection (Figure 2, right).

Reprint count factor f_{rc}, where RC is the percentage of novels with a low reprint count in the collection. There should be at least 30% of such novels, ideally 40-59%, but not more than 70%. f_{rc} has the highest value 11 when $RC \in [40, 60)$. Its value for SrpELTeC is 10, since there are 62 novels with a low reprint count (Figure 3, left).

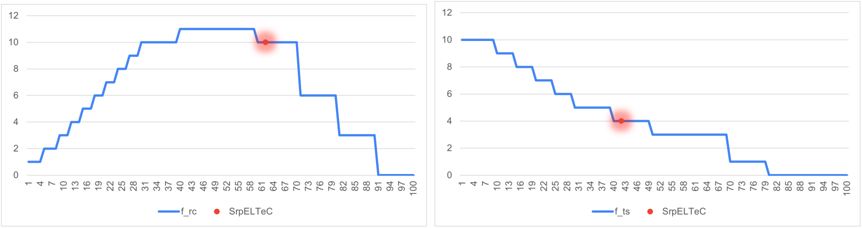


Figure 3. Reprint count factor $f_{rc} = f(RC)$ (left); Time slot factor $f_{ts} = f(TS)$ (right). Red dots are factor values for SrpELTeC.

Time slot factor f_{ts}, where TS is the range of the proportions of novels in each time slot. It is calculated as the difference between the highest and the lowest percentage of novels in corresponding time groups. This factor has the highest value 10, when the difference between percentages of novels in these time slot groups is less than 10. For SrpELTeC $f_{ts} = 4$ because there are 43 novels in T3 group and 2 novels in T1, thus the difference is 41 (Figure 3, right).

Size category factor for short novels f_{scs}, where SCS is the percentage of short novels in the collection. There should be at least 20% of short novels in a collection, ideally 33%, but not more than 60%. This factor has the highest value 11, when $SCS \in [30, 36]$. For SrpELTeC $f_{scs} = 10$ because there are 58 short novels (Figure 4).

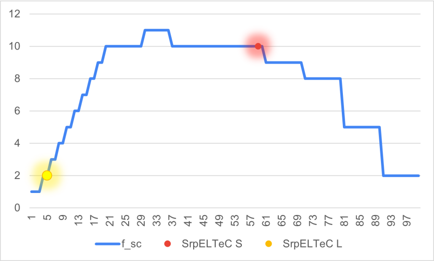


Figure 4. Size category factors: $f_{scs} = f(SCS)$ for short novels, $f_{scl} = f(SCL)$ for long novels. The red dot is the factor value for short novels, the yellow dot is the factor value for long novels for SrpELTeC.

Size category factor for long novels f_{scl}, where SCL is the percentage of long novels in the collection. There should be at least 20% of long novels in a collection, ideally 33%, but not more than 60%. This factor has the highest value 11, when $SCL \in [30, 36]$. For SrpELTeC $f_{scs} = 2$ because there are only 5 long novels (Figure 4).

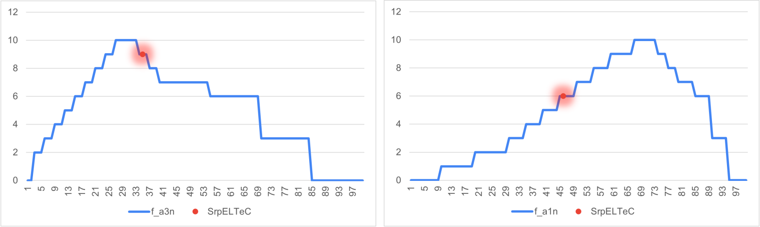


Figure 5. Three novels per author $f_{a3n} = f(A3N)$ (left); One novel per author $f_{a1n} = f(A1N)$ (right). Red dots are factor values for SrpELTeC.

Three novels per author factor, f_{a3n}, where $A3N$ is the percentage of novels written by authors represented with exactly 3 novels in a sub-collection. There should be at least 27% such novels, and no more than 33%. This factor has the highest value 10, when $A3N \in [27, 33]$. For SrpELTeC $f_{a3n} = 9$ because there are 12 authors represented with 3 novels (Figure 5, left).

One novel per author factor, f_{a1n}, where $A1N$ is the percentage of novels whose authors are represented in the collection by that novel only. There should be at least 67% such novels, and no more than 74%. This factor has the highest value 10, when $A1N \in [67, 73]$. For SrpELTeC $f_{a3n} = 6$ because there are 46 authors represented with exactly one novel (Figure 5, right).

The overall measure $E5C$ takes into account all these factors, with different weights according to their importance: collection size factor f_{cs}, being the most important, has the weight 3, author gender f_{ag}, reprint count f_{rc}, and time slot factor f_{ts} have the weight 2, while other factors' weight is one.

$$E5C = \frac{(f_{cs} * 3 + f_{ag} * 2 + f_{rc} * 2 + f_{ts} * 2 + f_{scs} + f_{scl} + f_{a3n} + f_{a1n}) * 10}{13} \tag{1}$$

A collection that is perfectly compliant to all balance criteria would thus have the highest values for each factor, and its $E5C$ would be:

$$E5C = \frac{(10 * 3 + 11 * 2 + 11 * 2 + 10 * 2 + 11 + 11 + 10 + 10) * 10}{13} = 104.6 \tag{2}$$

$E5C$ for SrpELTeC is:

$$E5C = \frac{(10 * 3 + 8 * 2 + 10 * 2 + 4 * 2 + 10 + 2 + 10 + 6) * 10}{13} = 78.46 \tag{3}$$

The Serbian sub-collection has the highest value of only one factor – the size collection $f_{sc} = 10$. The time slot factor has a rather low value, $f_{ts} = 4$ out of 10. This is due to the fact that there are only two novels for period the 1840-1859. It is interesting to note that there are actually some more Serbian novels written in that period, but they used the old orthography (before Karadžić's reform), were not modernized later, and as they are incomprehensible to contemporary readers and cannot be processed with tools for processing modern Serbian language, they ware not included.[14] A similar under-representation of novels in time slot T1 can be found in several other 100-novels sub-collections: Czech, Polish, Portuguese, Romanian, while only the Slovene sub-collection has as little as 2 novels, the same as the Serbian. Also, the Polish and the Slovene sub-collections have less novels in time slot T2 than Serbian, 11 and 13 respectively.

Another factor with a low value is the size category for long novels, $f_{scl} = 2$ out of 10. As explained before, Serbian authors tended to write, especially between 1840-1880, longer stories and novellas rather than novels. Also, the size of a novel measured by the number of words is a rather formal criterion that does not take into account the fact that some languages, like Serbian, are more "economical" in the use of words than others.

Presently, there are 10 sub-collections with 100 novels in ELTeC,[15] with $E5C$ ranging from 78.46 (Slovenian and Serbian) to 101.54 (French), with only three having $E5C$ close to 100.00 (English, French, and Hungarian). The fact that even these three sub-collections have not reached the highest $E5C$ shows how difficult that is. However, one should keep in mind that this measure was not developed in order to give gold, silver and bronze medals to the best scoring sub-collections. It was meant to indicate to future users of a sub-collection and the collection as a whole to what extent a specific sub-collection met the balancing criteria.

The value of $E5C$ significantly less than 100 for a 100 novel sub-collection can be the result of different circumstances, one of which is that for a certain language, due to the literary history of that language, there are not enough novels to fulfil all these complex criteria, e.g. not enough female authors, not enough long novels, etc. It is our firm belief that this is the case for Serbian.

6 Conclusion

In addition to the unquestionable cultural significance, the construction of this corpus will increase the visibility of Serbian literature in the world. The widest population will be offered a corpus that provides an insight into the development of the Serbian novel, revealing some little-known or forgotten

[14] One such novel is *Венацъ искренне любови Светоміра и Зорице : романтическа повѣ стъ сочинѣна Димитріемъ Михаиловићемъер* (Wreath of true love between Svetomir and Zorica : romantic story composed by Dimitrije Maihailović – free title translation) from 1840.

[15] Actually, two collections, French and Romanian, have 101 novels.

writers from the end of the second half of the 19th century and the beginning of the 20th century. This corpus will serve as the basis for the diachronic corpus of the modern Serbian language, which will be of great importance for studies of Serbian. The project and its outcome is also of immediate use for the modernization of Serbian lexicography, offering lexicographers text concordances and various ways to search them, as well as, in parallel, images of the original text.

This collection can be seen as a cornerstone for a future corpus that would cover not only other time periods but also other literary materials and represent a kind of cultural bridge to the synchronic and diachronic level of Serbian culture. This corpus will be offered in a modern digital environment to all interested researchers, who will have the opportunity to view and study this material with cutting-edge digital tools. It will also enable the creation of dictionaries of individual writers, which are lacking in the Serbian cultural scene.

Acknowledgment

We are greatly indebted to libraries that provided hard copies of novels and scanned them: University Library “Svetozar Marković”, National Library of Serbia and Matica Srpska Library. We are also grateful to numerous readers-volunteers recruited through the Society for Language Resources and Technologies (jerteh.rs) for their immense help in correcting and annotating novels.

We would like to thank Lou Burnard and other colleagues from the CA 16204 COST action, for helping in the various stages of SrpELTeC preparation.

References

Bacotić, Gordana, and Biljana Ristović. 2020. “Korisnici u COBISS okruženju.” *Organizacija Znanja – OZ* 25 (1–2): 1–11. https://doi.org/10.3359/oz2025004.

Juvan, Marko. 2019. *Worlding a Peripheral Literature (Canon and World Literature).* Palgrave Macmillan.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana. 2021. "The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata." *Infotheca - Journal for Digital Humanities* 21 (2): 26–42. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.2.

Krstev, Cvetana, and Ranka Stanković. 2020. "Old or new, we repair, adjust and alter (texts)." *Infotheca - Journal for Digital Humanities* 19 (2): 61–80. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2019.19.2.3.

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. "Annotation of the Serbian ELTeC Collection." *Infotheca - Journal for Digital Humanities* 21 (2): 43–59. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.3.

Stanković, Ranka, Mihailo Škorić, and Petar Popović. 2021. "SrpELTeC on platforms: *Udaljeno čitanje*, Aurora, noSketch." *Infotheca - Journal for Digital Humanities* 21 (2): 136–153. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.7.

Stevenson, Jay. 2007. *The Complete Idiot's Guide to English Literature.* Alpha Books (Penguin Group).

TEI Consortium. 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.3.0.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Деретић, Јован. 1981. *Српски роман: 1800-1950.* Нолит.

Деретић, Јован. 1983. *Историја српске књижевности.* Нолит.

214Тртовац, Александра, Јелена Андоновски, and Наташа Дакић. 2021. "Дигитална библиотека Универзитетске библиотеке "Светозар Марковић" – од скенираних страница до претраживе колекције." *Библиотекар: Орган Друштва Библиотекара НР Србије* 63 (1): 27–48.

The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata

UDC 004.65:02

DOI 10.18485/infotheca.2021.21.2.2

Cvetana Krstev
cvetana@matf.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

ABSTRACT: In this paper we present the metadata assigned to the Serbian ELTeC sub-collection, analyse them and draw some conclusions about titling practices of Serbian narrative literature in the period 1840-1920, authors' gender and age, publication places, modes of publication, as well as about institutions and individuals that are due credits for the production of SrpELTeC.

KEYWORDS: metadata, TEI header, Serbian Literature, ELTeC corpus.

PAPER SUBMITTED: 05 November 2021
PAPER ACCEPTED: 14 November 2021

1 Introduction

The practice of assigning descriptive data to some piece of information has a long history, ever since librarians started to produce catalogues. According to Hodge (2001, 3) "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. Metadata is often called data about data or information about information". These descriptive data was not always called "metadata", but, as pointed by (Duval et al. 2002): "The rapid changes in the means of information access occasioned by the emergence of the World Wide Web have spawned an upheaval in the means of describing and managing information resources. Metadata is a primary tool in this work, and an important link in the value chain of knowledge economies." Metadata assigned to information of any kind brings additional value to them: being well described, this information can be more easily retrieved by different users, for various purposes, over a long period of time.

The need to assign metadata to literary works in digital form was recognized many years before the WWW upheaval mentioned in (Duval et al. 2002). From the very beginning, and that means the first published guidelines in the early 1990s, the TEI prescribed, as mandatory, the use of headers for all TEI conforming documents (Burnard 2013). The TEI header was seen as the primary source of information about a digital source, corresponding to a certain extent to a cover page of a printed book, and consequently to a library catalogue record. In the following section, we will explain which metadata can be found in headers of literary works of SrpELTeC, and which aggregated information can be deduced from them.

2 Metadata Header

Since all texts in the ELTeC corpus are TEI documents, each of them has its own TEI header[1] – The ELTeC Header.[2] The content of the header was agreed by the action's Working Group 1 – Scholarly Resources/ELTeC (WG1) with the aim of making it informative, but at the same time simple.

Each ELTeC header has three obligatory elements:

- `<fileDesc>`, which describes the digital, that is ELTeC edition (element `<titleStmt>`), its size (element `<extent>`), availability and licensing (element `<publicationStmt>`), and source(s) from which it was derived (element `<sourceDesc>`);
- `<profileDesc>`, which gives additional information about the ELTeC edition, about the language the text was written in (element `<langUsage>`), and particularly about its characteristics that bear upon the balance of the whole sub-collection (element `<textDesc>`);
- `<revisionDesc>`, which records all changes of the document after its first publication.

The element `<sourceDesc>` contains one or more bibliographic descriptions (element `<bibl>`). The bibliographic description for the first edition is obligatory (attribute value `firstEdition`). However, if the first edition was not used to prepare the ELTeC edition, than additional `<bibl>` elements are necessary: `<sourceEdition>` for the print source or `<digitalSource>` for the digital source. Of course, more than two `<bibl>` elements can occur if, for instance, the ELTeC edition was derived from some digital edition that came from a print edition that was not the first one.

Here, we give, as an example, the ELTeC header of the novel *Jarani* (Friends) (SRP19132) from SrpELTeC; some lines are shortened, and some elements deleted to make it more readable.

[1] About the TEI header in TEI P5 guidelines TEI header.

[2] About the ELTeC

```
<teiHeader>
<fileDesc>
  <titleStmt>
    <title>Јарани : приповетка : ELTeC издање</title>
    <author ref="viaf:136335">
       Ћоровић, Светозар (1875-1919)
    </author>
    <respStmt>
      <resp>Скенирање</resp>
      <name>
        Универзитетска библиотека "Светозар Марковић"
      </name>
    </respStmt>
    <respStmt>
      <resp>OCR и корекција текста</resp>
      <name>Тања Ћулафић</name>
      <name>Цветана Крстев</name>
    </respStmt>
    <respStmt>
      <resp>Кодирање за ELTeC</resp>
      <name>Цветана Крстев</name>
    </respStmt>
  </titleStmt>
  <extent>
    <measure unit="words">47623</measure>
    <measure unit="pages">199</measure>
  </extent>
  <publicationStmt>
    <publisher ref="https://distant-reading.net">
      COST Action "Distant Reading ..." (CA16204)
    </publisher>
    <distributor ref="https://zenodo.org/communities/eltec/">
    Zenodo.org
    </distributor>
    <date when="2021-04-09"/>
    <availability> <!-- about licence --></availability>
    <!-- references to various releases -->
  </publicationStmt>
  <sourceDesc>
```

```
    <bibl type="printSource">
      <author>Ћоровић, Светозар</author>
      <title>Дјела. Књ. 2</title>
      <publisher>Свјетлост</publisher>
      <pubPlace>Сарајево</pubPlace>
      <date>1953</date>
      <idno type="COBISS-SR.ID">54425863</idno>
    </bibl>
    <bibl type="firstEdition">
      <author>Ћоровић, Светозар</author>
      <title>Јарани : приповетка</title>
      <pubPlace>Нови Сад</pubPlace>
      <publisher>Матица српска</publisher>
      <date>1913</date>
      <idno type="COBISS-SR.ID">16904711</idno>
    </bibl>
  </sourceDesc>
</fileDesc>
<profileDesc>
  <langUsage>
    <language ident="sr-Cyrl" usage="100">српски</language>
  </langUsage>
  <textDesc>
    <authorGender xmlns="http:...ns" key="M"/>
    <size xmlns="http:...ns" key="short"/>
    <reprintCount xmlns="http:...ns" key="low"/>
    <timeSlot xmlns="...ns" key="T4"/>
  </textDesc>
</profileDesc>
<revisionDesc>
  <change when="2021-04-09">
    Converted by checkUp script for new release
  </change>
  <!-- other changes -->
</revisionDesc>
</teiHeader>
```

3 The Analysis of Data

Each novel in the SrpELTeC sub-collection has its own metadata header, as presented in the previous subsection. When preparing headers, care must be taken to use all its obligatory elements and fill in correct and consistent data. The use of elements and their sub-elements in the correct way is checked by the appropriate XML Schema, but the header producer has to take care about the content. For instance, the author's name should be given in the reverse order, and dates of her/his birth and death in parentheses.

As a result, we obtained consistent headers, one of which is presented in the previous subsection. Developers of most of other language sub-collections within ELTeC use similar headers, while some contain additional information. For instance, the headers of French sub-collection use the element `<textClass>` (within `<profileDesc>`) to indicate the narrative perspective and subgenre. The headers of Slovene sub-collection use the forth sub-element of a TEI header, `<encodingDesc >`, to describe briefly the process of transformation of existing digital edition to the ELTeC edition.

In the following subsections we will present some conclusions that can be derived about the SrpELTeC sub-collection from metadata assigned to each novel.

3.1 Titles

Titles of novels in the SrpELTeC sub-collection were taken from the first editions, whenever possible.[3] It was noted that the titles of subsequent editions sometimes deviate from first editions. For instance, the novel *Seljanka* by Janko Veselinović (SRP18932) was first published in instalments in 1884 in the journal *Otadžbina* as *Seljanka. Roman* (The peasant woman. Novel), its first book edition in 1893 (used for scanning) was entitled *Seljanka – pripovetka iz seoskog života* (The peasant woman – a story from rural life). The subsequent editions used both variants as well as simply – *Seljanka.* In rare occasions the first edition titles had to be adapted to modern script, as was the case with the novel *Đurađ Branković* by Jakov Ignjatović (SRP18590) that still did not use the reformed alphabet: *Ђурађъ Бранковићъ : (историчный романъ).*

In paper (Patras et al. 2021) authors analysed titles in eleven ELTeC sub-collections, most of which were not complete at the time the article was

[3] If the first edition was scanned, then the title was taken as stated on the front page; otherwise we relied on the library catalogue.

written. The Serbian sub-collection contained 50 novels and some observations were:[4]

- Titles of Serbian novels used only location entities that could be traced on the map, e.g. *Novac: roman iz beogradskog života* (SRP1906), which was not the case for other language sub-collections;
- Serbian novels had more "long" titles (6 words or more) than other sub-collections, approximately 25%;
- In Serbian novels, as well as Slovenian and Ukrainian, the existence of persons occurring in titles is more often than not impenetrable, meaning that their "roles" in novels have to be discovered by readers, e.g. *Milan Narandžić* (SRP18630) and *Radetića Mara: pripovetka iz seoskog života* (Radetića Mara: a story from rural life) (SRP18940).

While it would be interesting to extend this research on titles to the complete ELTeC collection, we will make here only some observations. The average length of titles in the 100-novels Serbian collection is 5.19 words, with only 8 one-word titles, e.g. *Sanjalo* (Dreamer) (SRP18880), and as many as 40 novels having titles of 6 or more words. The longest title has 18 words *Zločin jedne svekrve : kriminalna pripovetka : iz skore prošlosti / po autentičnim podatcima i sudskim aktima napisao K. D. Jezdić* (The crime of a mother in law : a crime story : from the recent past : written on the basis of authentic data and court acts by K. D. Jezdić) (SRP19062). This title also has as many as three subtitles. Only 40 novels from the whole sub-collection do not have subtitles, 52 have one subtitle, and 7 two.

The subtitles usually contain genre indicators – 53 out of 60 subtitles – and these are: *roman* (novel) 25, *pripovetka/pripovijetka* (short story) 21, *novela* (novella) 3, *slika/slike* (sketches) 3, *crta* (outline) 1. Although one would expect that all indicators except *roman* point to shorter narratives of a fragmented nature it is not always the case. As we already saw, the indicator of *Seljanka* (SRP18932) changed from one edition to another. Also, the indicator for *Pop Ćira i Pop Spira* (Father Ćira and father Spira) (SRP18941) in the first edition is short story (*pripovetka*), while today nobody would consider this work anything but a novel (*roman*).

3.2 Authors

The works represented in the SrpELTeC subcollection are written by 66 different authors: one author is represented with 4 novels (Jakov Ignjatović),

[4] The whole data set and visualizations used in (Patras et al. 2021) can be found in (Patras et al. 2020).

12 with 3 novels (Svetozar Ćorović, Jelena Dimitrijević, Vladan Đorđević, Draga Gavrilović, Lazar Komarčić, Čedomilj Mijatović, Milan Đ. Milićević, Svetolik Ranković, Stevan Sremac, Borisav Stanković, Milutin Uskoković, Janko Veselinović), 7 with 2 novels (Andra Gavrilović, Dragutin J. Ilić, Đura Jakšić, Tadija P. Kostić, Branislav Nušić, Pera Todorović, Ivo Ćipiko), and 46 with one novel (see Figure 1. This distribution does not fit perfectly into the criteria for balanced representation of authors; the reasons for that are discussed in paper (Trtovac, Milnović, and Krstev 2021) of this issue.[5]

In the SrpELTeC sub-collection there are no works with common authorship of several authors, while the author of one work is unknown – *Beogradske tajne* (Belgrade's secrets) (SRP18923). The author's name is not given in the printed edition, neither was his/her name identified later. Only 8 novels are written by (4 different) "authoresses": Draga Gavrilović, Jelena Dimitrijević, Milica Janković and Isidora Sekulić. When compared with other ELTeC collections, which are complete comprising 100 works, one can observe that in all of them women are better represented: while English collection has a perfect balance (51 female authors), and German, French and Polish have a high representation of female authors (33, 34, 42, respectively), in other collections female authors are not so well represented, but nevertheless, better than in SrpELTeC: in Czech 12, Hungarian 21, Portuguese 17, Romanian 16, and Slovenian 11.

Twelve authors (not counting one unknown author), out of 66 different authors, are not represented in VIAF.[6] The authors' birth and death years were taken form the library catalogues, and where they were missing, we tried to retrieve them from some older encyclopedias (Петровић 1937; Бихаљи-Мерин 1959a, 1959b); they were, however, rarely found. Two authors with unknown birth and death dates are recorded in VIAF (Stevan Mamuzić and Dimitrije Tasić), while this data is known for three authors with no record in this database: Panta Popović (1843-1918), Živojin Jovičić (1837-1908), Boško St. Petrović (1869-1913). For one author recorded in VIAF, Dušan Rogić, the year of birth is known (1855) while the year of death remains unknown.

The author who was born the earliest, in 1817, was Jovan Subotić, while the author who was born the latest was Mladen St. Đuričić, in 1889. The

[5] The final edition of SrpELTeC differs slightly from the one presented here; data about novels and authors in SrpELTeC at the time this article was written as well as in the final edition can be found in (Krstev and Stanković 2021) in this issue.

[6] The Virtual International Authority File

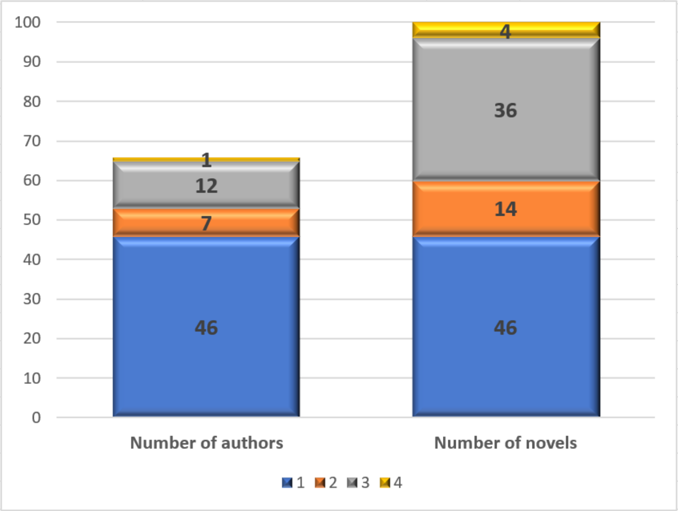


Figure 1. (a) Number of novels that are represented in SrpELTeC corpus by one to four novels; (b) number of novels by these authors in SrpELTeC.

author who died the earliest, in 1859, was Bogoboj Atanacković, while the author who died the latest is again Mladen St. Đuričić in 1987.[7] Milutin Uskoković, the author of three novels in SrpELTeC, lived the shortest – 32 years, while, as expected, Mladen St. Đuričić lived the longest – 99 years. The average life length of authors was 61.5 years.

The novel written at the youngest age of an author is *Kočina krajna* (Koča's frontier) by Vladan Đorđević (SRP18631), who was 20 years old at the time of the novel's publication. Other very young authors were Dragutin Ranković, who wrote the novel *Slavko* (SRP19023) at the age of 21, and Branko Mihajlović who wrote *Pred zoru* (Before dawn) at the age of 22. Stojan Novaković was the oldest when he published the novel *Kaluđer i hajduk : pripovetka o poslednjim danima Srbije u XV veku* (A monk and a

[7] It is interesting to note that Mladen St. Đuričić wrote his novel, or rather novella *Kad šume talasi* (When the waves rustle) (SRP19141), in 1914, when he was 25 years old. He was a captain of river navigation – like the main character of his novella – and besides poems, stories, and essays, also wrote professional books, most notably *Istorija jugoslovenskog rečnog brodarstva* (History of Yugoslav river shipping) in 1965.

haiduk : a story about the last days of Serbia in 15$^{\text{th}}$ century) (SRP18730) – he was 72 years old. On the average, authors were 38.8 years old when their selected works were published for the first time.

3.3 The Year of the First Publication

As the year of the first publication of novels we have chosen the year in which the work was first published either as a book (83), or in instalments in literary journals (16). If the publication in instalments span over more than one year, we have taken as the year of publication the year when the last instalment was published. To obtain this information, we relied primarily on the OPAC library catalog, encyclopedias (Петровић 1937; Бихаљи-Мерин 1959a, 1959b), reference books (Деретић 1981, 1983; Милисавац 1972), and sometimes on the reference editions of a particular author's works. For the novel *Seljaci* (Peasants) by Đura Jakšić (SRP1974) we could not determine the mode of its first publication, but found the year of publication in the collected works of this famous author (Јакшић 1883).

The distribution of publications in instalments over time slots is as follows: one in T1 (1840-1859), i.e. 50.0% of all novels from that time slot, 7 in T2 (1860-1879), also 50.0%, 5 in T3 (1880-1899), i.e. 11.4%, and 2 in T4 (1900-1920), i.e. 5.0%. This data show that publication in instalments was losing popularity over time.

It is obvious that all years from time slots T1 and T2 do not have their representatives in SrpELTeC, since there are more years than published works for these slots. In time slot T3, in all years save three – 1883, 1885 and 1890 – one or several works were published that are represented in SrpELTeC. The situation in time slot T4 seems similar, as there are three years with no representative in the Serbian sub-collection: 1915, 1916, 1917, years that coincide with the World War I. That this was not the time for novel writing and publishing in Serbia is supported by the fact that these years of publishing do not occur neither among 11 novels of the extended sub-collection SrpELteC-ext nor among 39 novels waiting to be prepared for this sub-collection. In other 100-novel ELTeC sub-collections, one cannot observe something similar: Czech and Slovene have no gap between years 1914 and 1918, while for other sub-collections, one or two years are not represented. However, given the large literary production in these languages, this could be for some other reason than war.

3.4 Publishers and Places of Publication

As pointed in the previous subsection, 15 novels were published for the first time in the following journals:

- *Javor* (Јовановић Змај 1862–1893) (4);
- *Otadžbina* (Ђорђевић 1875–1892) (4);
- *Сербскій лѣтописъ* (Субботићъ 1842–1855) (1);
- *Danica* (Рајковић 1860–1872) (1);
- *Matica* (Хаџић 1865–1870) (1);
- *Nova iskra* (Одавић 1899–1911) (1);
- *Srpski književni glasnik* (Поповић 1901–1941) (1);
- *Stražilovo* (Грчић 1885–894) (1);
- *Male novine* (Кимпановић 1888–1903) (1).[8]

The publisher that has published the most novels in book form – *Srpska književna zadruga*, with 11 novels, as well as publishers *Zadužbina Ilije M. Kolarca*, with 3 novels, and *Matica srpska* with 2, have all survived to the present day. The remaining publishers *Matica hrvatska*, *Književni jug*, *Srpski pregled*, *Braća Savić*, *Savić i komp.* are represented with only one novel in SrpELTeC.

In the time period covered by ELTeC the roles of publishers, printers and bookstores were not so clear-cut as today. Thus, besides 21 publishers mentioned before, as many as 42 novels, on their cover pages,[9] instead of a publisher, mention a printing house (*štamparija*), a total of 32 different, e.g. *Državna štamparija kraljevine Srbije* (State Printing House of the Kingdom of Serbia) or *Parna štamparija D. Dimitrijevića* (Steam printing house of D. Dimitrijević). In other cases bookstores are mentioned (5), e.g. *Knjižara Velimira Valožića* (Bookstore of Velimir Valožić), political parties (1) *Napredna stranka* (Progressive Party), or persons (11), where in one case the person mentioned is the author himself – *Bespuće* by Veljko M. Milićević (SRP19121). The publisher is unknown in two cases.

[8] The only information about the publisher of *Beogradske tajne* (SRP18923) on the book cover and in the library catalog is the name *Zabavnik Malih novina* (The Entertainer of the Little Newspaper). It is not clear whether *Zabavnik* (The Entertainer) was related to the daily newspaper *Male novine* (Little Newspaper), although an exhaustive note in the catalog record for *Male novine* mentions that the subtitle for 1893 was “Dnevni list za svakoga sa zabavnim mesečnim dodatkom” (A daily newspaper for everyone, with a monthly entertainment supplement). Since the edition has the form of a book, we treated it as such.

[9] In cases the first editions were not available we relied on the catalog data.

Most of the novels were published in Belgrade (59), Novi Sad (18) and Sremski Karlovci (4). One novel was published in each of the following places, which are within today’s borders of Serbia: Veliki Bečkerek (today Zrenjanin), Vršac, Kikinda, Kragujevac, Mitrovica (most probably Sremska Mitrovica), and Niš. Some novels were published outside today’s borders of Serbia: Zagreb (3), Sarajevo (2), Mostar (2), and one novel in Beč (Vienna), Budim (the western part of today’s Budapest), and Dubrovnik. For three novels the place of publication is unknown (see Figure 2).

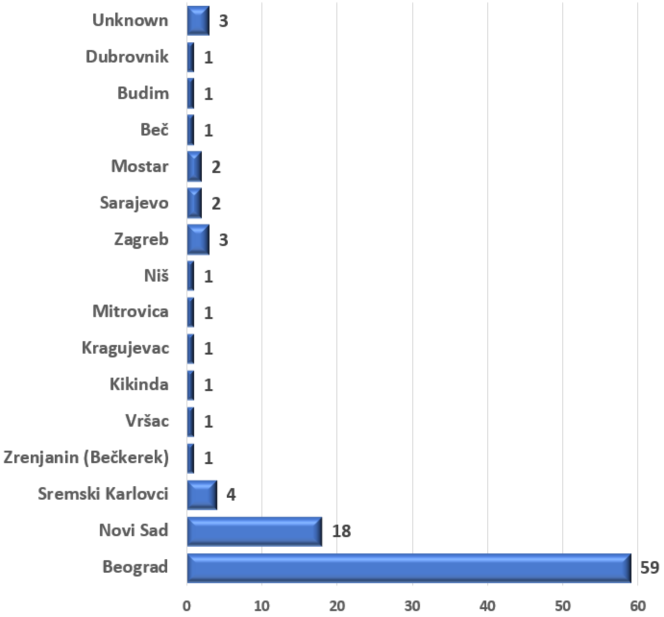


Figure 2. Publications per places of publication

3.5 Language and Alphabets

All novels are written in the Serbian language, although short passages in other languages appear almost in all of them, and are annotated with the

`<foreign>` tag. In only one novel, *Nove* (New women) (SRP19120), the passages in French and English appear more often and were roughly estimated to 1% of the whole text. This information is contained in the header.

All sources except two, whether printed or digital, used Cyrillic script. Two novels, *Pre sreće* (Prior to happiness) (SRP19180) and *U ćelijama* (In the cells) (SRP19192), both published in Zagreb, used Latin script.[10] The script of the sources was preserved in the electronic edition.

The Cyrillic alphabet in printed sources is the Serbian Cyrillic, as developed at the beginning of the 19th century by Vuk Stefanović Karadžić. Only the initial publication of the novel *Kaluđer : istina i poezija* (The monk: truth and poetry) (SRP1873), from the year 1881, used the Cyrillic letter Yat, which Vuk removed from the alphabet. Since this novel was never republished, we used this initial publication for SrpELTeC and kept the letter Yat.[11] Several other novels used occasionally the soft sign (Ь), which was also removed from the alphabet or merged with other letters; we kept it as well. The use of these letters in the text is not recorded in the header.

3.6 Sources and Scanning

As explained in (Trtovac, Milnović, and Krstev 2021), we wanted to use the first editions for the SrpELTeC sub-collections whenever possible. This was, however, impossible for 15 novels, first published in journals, because not all journals (or issues) were available, or if they were already digitized, they were not suitable for further processing, as the OCR on scanned journals in general produces very poor results. In certain cases the first editions of books could not be used either, because they were not available or the OCR also produced very poor results. This was the case with *Opštinsko dete* (SRP19021), for which the first edition was retrieved and scanned, the OCR was done, but the result was so bad that we had to repeat the whole process with a later edition. As this was a very time consuming process, in four cases we decided to use already digitized novels, despite the fact that the solution was not optimal, as explained in (Trtovac, Milnović, and Krstev 2021).

[10] One more novella used the Latin script for its first edition published in Belgrade in 1920: *Put Alije Gjerzeleza* (The Journey of Alija Gjerzelez) by Ivo Andrić, the only Serbian and Yugoslav Nobel Prize Winner for Literature. Unfortunately, it is shorter than 10K words and was included in SrpELTeC-ext.

[11] We are particularly grateful to Aleksandra Davidović who read and corrected this novel, which was a very demanding task.

As a result, SrpELTeC comprises 56 novels scanned from the first editions, 40 novels scanned from a later edition and 4 novels that were already digitized. When we used later editions for scanning, we tried to use the edition that was time-wise the closest to the first edition. SrpELTeC comprises 14 later editions of novels published less than 10 years after the first edition, 5 later editions published less than 20 years after the first, and 2 later editions published less than 50 years after the first edition. Eleven later editions scanned for SrpELTeC were published between 50 and 100 years after the first edition, and 6 later editions more than 100 years after the first. For the novel *Nazareni* (Nazareans) (SRP18966) we used the second edition, which appeared the same year as the first, while the edition used for the novel *Đurađ Branković* (SRP18590) was published 129 years after the first. On the average, the print editions used for scanning that were not the first editions were published 34.8 years after the first.

The novels were procured from several libraries, but by far the most from the University Library "Svetozar Marković" – 68. The other libraries were: the National Library of Serbia (8), the Library of "Matica srpska" (3), the Library of the Serbian Academy of Sciences and Art (2), and the private library of Duško Vitas and Cvetana Krstev (12). Four already digitized novels were downloaded from the digital library *Antologija srpske književnosti* (The Anthology of Serbian literature), maintained by the Teacher Education Faculty at the University of Belgrade.[12] Two novels were retyped: *Đul-Marikina prikažnja* (Đul-Marika's narration) (SRP19012)[13] and *Došljaci* (Newcomers) (SRP19100).[14]

All novels were corrected and annotated by a number of volunteers, in line with the recommendations of the action's WG1 (as explained in (Trtovac, Milnović, and Krstev 2021)).[15] All volunteers and their contribution to the preparation of SrELTeC are listed in Table 1.[16] The contribution of readers is

[12] Antologija srpske književnosti

[13] The first edition of the novel could not be procured, so it was retyped from its scanned version available at the National Library of Kruševac Literary opus of Jelena Dimitrijević

[14] This novel was previously retyped by students of Library and Information Sciences at the University of Belgrade, Faculty of Philology as part of their practical assignments. It was further checked, corrected and prepared for SrpELTeC.

[15] The volunteers were mainly members or supporters of the Society for Language Resources and Technologies JeRTeH, as well as MA and PhD students at the University of Belgrade, for whom this was a part of their practical assignment.

[16] It should be noted that some volunteers read additional novels, which were transferred to SrpELTeC-ext: Cvetana Krstev +4 (words 53,218; pages 237), Ivan

quantified here by the number of novels read, and the total number of words and printed pages. However, one important parameter is missing, and that is the quality of the text after OCR and automatic correction (see (Trtovac, Milnović, and Krstev 2021)). Namely, some shorter texts can be much harder to work on than longer texts, but we did not quantify this parameter.

Reader	**bks**	**words**	**pages**
Cvetana Krstev	32	1,669,442	7,572
Duško Vitas	13	607,156	2,708
Ivan Obradović	9	592,497	2,534
Ranka Stanković	7	255,718	1,192
B. Šandrih	4	189,575	805
Olivera Kitanović	4	130,493	559
Biljana Lazić	3	120,890	549
A. Tomašević	3	111,913	593
A. Davidović	2	113,385	517
J. Dimitrijević	2	92,327	361
Tanja Ćulafić	2	68,481	296
J. Andonovski	2	60,262	213
Miloš Utvić	1	81,827	249
Milena Mihajlović*	1	71,824	400
M. Ikonić Nešić	1	61,971	271
Nenad Zekavica*	1	60,464	296
A. Marković	1	59,480	204
Đorđe Stakić	1	51,288	276
Anica Milanović*	1	44,760	125
Andrea Adamović*	1	37,860	181
Milica Antić*	1	29,543	125
Vanja Radulović*	1	20,414	92
Tamara Radak	1	14,247	52
Stefan Stepanović	1	12,023	36
A. Jovanović	1	10,095	80
total	96	4,791,090	20,288

Table 1. List of readers that contributed to the preparation of the SrpELTeC. Master and PhD students are marked with an *, the others are JeRTeH volunteers.

4 Conclusions

In this paper we presented the metadata assigned to each novel of the ELTeC corpus, with the emphasis on the SrpELTeC sub-collection. From these data we could draw some conclusions about titling practices of Serbian narrative literature in the period 1840-1920, authors' gender and age, publication places, modes of publication, as well as about institutions and individuals that are due credits for the production of SrpELTeC.

We also believe that future users of ELTeC corpus, and SrpELTeC as its part, will benefit from the assigned metadata. We also hope that these users

Obradović +4 (words 81,227; 487), Duško Vitas +1 (words 20,286; pages 100), Stefan Stepanović +1 (words 7,105; pages 41), Sergej Adamov +1 (words 22,904; pages 122).

will enrich metadata with information related to their expertise. One can note that information about dialect and pronunciation used in novels, origin of authors, as well as novels’ genre, are still missing, and should be added by experts.

References

Burnard, Lou. 2013. “The evolution of the Text Encoding Initiative: from research project to research infrastructure.” *Journal of the Text Encoding Initiative,* no. 5.

Duval, Erik, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. 2002. “Metadata principles and practicalities.” *D-lib Magazine* 8 (4): 1–10.

Hodge, Gail M. 2001. *Metadata made simpler.*

Krstev, Cvetana, and Ranka Stanković. 2021. “Novels and Authors of the Serbian ELTeC Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 172–186. ISSN: 2217-9461.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2020. *Dataset for ELTeC titles.* Zenodo. https://doi.org/10.5281/zenodo.4268669.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2021. “Thresholds to the “Great Unread”: Titling Practices in Eleven ELTeC Collections.” *Interférences littéraires/Literaire interferenties* 25:163–187.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. “The Serbian Part of the ELTeC Collection – from the Empty List to the 100 Novels Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.1.

Бихаљи-Мерин, Ото и др., ed. 1959a. *Мала енциклопедија Просвета: општа енциклопедија. 1, А - Љутомер.* Тираж 30000. Просвета.

Бихаљи-Мерин, Ото и др., ed. 1959b. *Мала енциклопедија Просвета: општа енциклопедија. 2, М - Шчукин.* Просвета.

Грчић, Јован, ed. 1885–894. *Стражилово : лист за забаву, поуку и уметност.* Дружина око “Стражилова”. http://ubsm.bg.ac.rs/cirilica/zbirka/novina/strazilovo-1885-1894.

Деретић, Јован. 1981. *Српски роман: 1800-1950.* Нолит.

Деретић, Јован. 1983. *Историја српске књижевности.* Нолит.

Ђорђевић, Владан, ed. 1875–1892. *Отаџбина : књижевност, наука, друштвени живот.* Владан Ђорђевић. http://ubsm.bg.ac.rs/cirilica/zbirka/novina/otadzbina-1875-1892.

Јакшић, Ђура. 1883. *Дела Ђуре Јакшића. Књ. 5, Приповетке.* У краљевско-српској држ. штампарији.

Јовановић Змај, Јован, ed. 1862–1893. *Јавор: лист за забаву и науку.* Јован Јовановић Змај. http://www.digitalna.nb.rs/sf/NBS/casopisi_pretrazivi_po_datumu/P_247%20http://digital.bms.rs/pub.php?s=RPSr-III-15.

Кимпановић, Ђорђе, ed. 1888–1903. *Мале новине : дневни лист за свакога.* Пера Тодоровић. http://ubsm.bg.ac.rs/cirilica/zbirka/novina/male-novine-1888-1903.

Милисавац, Живан, ed. 1972. *Приповедачи.* Vol. 66. Матица српска; Српска књижевна задруга.

Одавић, Риста Ј., ed. 1899–1911. *Нова Искра : илустровани лист.* Р. Ј. Одавић. http://ubsm.bg.ac.rs/cirilica/zbirka/novina/iskra-1899-1911.

Петровић, Петар М. и др., ed. 1937. *Свезнање: општи енциклопедиски лексикон у једној књизи.* Народно дело.

Поповић, Богдан, ed. 1901–1941. *Српски књижевни гласник.* Светислав С. Симић. http://www.digitalna.nb.rs/wb/NBS/casopisi_pretrazivi_po_datumu/Srpski_knjizevni_glasnik.

Рајковић, Ђорђе, ed. 1860–1872. *Даница : лист за забаву и књижевност.* Ђорђе Поповић. http://www.digitalna.nb.rs/sf/NBS/casopisi_pretrazivi_po_datumu/P_0151%20http://digital.bms.rs/pub.php?s=RPSr-II-15%20http://digital.nb.rs/scc/browse.php?collection=no-danica.

Субботићь, Іоаннъ, ed. 1842–1855. *Сербскій лѣтописъ*. Писмены Крал. Свеучилишта Пештанскогъ. http://digital.bms.rs/pub.php?s=RPSr-II-4.

Хаџић, Антоније, ed. 1865–1870. *Матица : лист за књижевност и забаву*. Матица србска. http://www.digitalna.nb.rs/sf/NBS/casopisi_pretrazivi_po_datumu/matica.

Scientific paper

Annotation of the Serbian ELTeC Collection

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2021.21.2.3

ABSTRACT: This paper presents the so-called level-2 edition of SrpELTeC collection developed within the activities of Working Group 2 - Methods and Tools of the COST Action CA 16204 (Distant Reading for European Literary History), and its schema specification. The level-2 edition is a follow-up of the level-1 edition, which is used as input for morphosyntactic and NER annotation of novels. The Serbian level-2 pipeline outlines steps required for production of level-2, including methods and tools used in the process. Some statistics drawn from the Serbian ELTeC level-2 sub-collection brings an interesting insight into collection content.

KEYWORDS: distant reading, literary corpus, tagging, NER, lemmatization, ELTeC.

PAPER SUBMITTED: 17 November 2021
PAPER ACCEPTED: 25 December 2021

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Geology and Mining
Belgrade, Serbia

Cvetana Krstev
cvetana@matf.bg.ac.rs
Branislava Šandrih Todorović
branislava.sandrih@fil.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Geology and Mining
Belgrade, Serbia

1 Introduction

Working group "Methods and Tools" (WG2) of the COST action "Distant Reading for European Literary History" (CA16204) is concerned with text analytic techniques and tools. WG2 coordinates activities related to sharing, evaluating, adaptation and improving methods and tools for Distant Reading research, and establishing best practices across Europe. It has a large range of activities, from the creation of manual reference annotations for the evaluation of automatic annotation tools, to the annotations' integration strategy. Namely, one of the problems WG2 tackled is integration of results of tokenization, lemmatization, part-of-speech tagging, and Named Entity Recognition (NER) into one document conforming to the ELTeC XML/TEI format. The existing tools are analysed and some guidelines for their application are published while others are still under development. Members of WG2 are active in development of NLP resources and tools, information extraction, computational linguistics, text mining, computational stylistics, and digital literary studies.

Two main problems encountered in producing Serbian ELTeC level-2 were similar to those encountered for other languages: 1) majority of morphosyntactic taggers do not work well with XML format and 2) harmonization of NER and morphosyntactic annotations, which are performed separately with different tools. A solution was found in the TXM tool[1] (Heiden 2010; Heiden, Magué, and Pincemin 2010), an environment that enables tagging of XML files, which solves the problem of alignment of NER and morphosyntactic tags. TXM also enables the construction of a sub-corpora or partitions based on metadata (date, author, genre, etc.) or corpus structural units (like text, chapter, paragraph), querying (using the CQP browser), and the processing of more complex query results using quantitative methods (supported by the R statistical package), as well as the export of results in a tabular or graphical form (Jaćimović 2019).

The second section of this paper "Level-2 specification" will introduce concepts and current state of the schema used for morphosyntactic and NER annotation of level-1 form of novels. The third section "Serbian level-2 pipeline" will introduce steps required for the production of level-2, including methods and tools used in the process. The fourth section "SrpELTeC level-2 statistics" will bring some numerical insights from the developed dataset.

2 Level-2 specification

The encoding of novels is produced in incremental levels, each validated by the appropriate RELAXNG schemas.[2] Description of level-2 schema is given in Encoding Guidelines for the ELTeC: level 2 (distantreading.github.io).[3] At the time of writing this paper, schema for level-2 was not yet finalized, but it is expected to be done soon. ELTeC level-2 includes all elements existing in level-1 and introduces some new ones: `<s>` as the sentence tag, used for segmentation of text into sentences, and `<w>` and `<pc>`, used for tokenization of text into tokens, and their annotation. Individual words are marked using the `<w>` element and mandatory linguistic attributes `@pos`, `@lemma`, and

[1] TXM is using the CQP (Corpus Query Processor) browser build on IMS Open Corpus Workbench and the R statistical package

[2] ELTeC Schemas

[3] Encoding Guidelines for the ELTeC: level 2

@join, as well as some optional attributes like the general XML attribute @xml:id for unique identification and @msd for more detailed morphosyntactic description. As tokens can be both words and punctuation marks, as well as other special characters, TEI recommends that these two cases should be distinguished by using two different elements: <w> for words and <pc> for punctuation and special characters.

The proposal is to eliminate any content within a <ref> element at level 2. The elements <p>, <head>, <note> and <l> can contain a sequence of <s> elements, while elements <gap>, <milestone>, <pb>, and <ref> are also permitted within text content at any point, but are disregarded in segmentation (Burnard, Schöch, and Odebrecht 2021). The element <s> can contain a sequence of <w> elements, either directly or in the sub-paragraph elements <corr>, <emph>, <foreign>, <hi>, <label>, <title>. The TEI element <rs> (referring string) has a special purpose in the level-2 format: it is used for the encoding of named entities, such as people, their roles, locations, organisations, works, events, and demonyms (Frontini et al. 2020; Šandrih Todorović et al. 2021).

WG2 had several physical meetings, first in Prague (Czech Republic), Antwerp (Belgium), Lisbon (Portugal), Budapest (Hungary) and in Malaga (Spain), and several online meetings for smaller teams focused on special topics, such as: morphosyntactic tagging, NER, direct speech, semantic analysis. Some resources developed by WG2 are available in the github repository.[4]

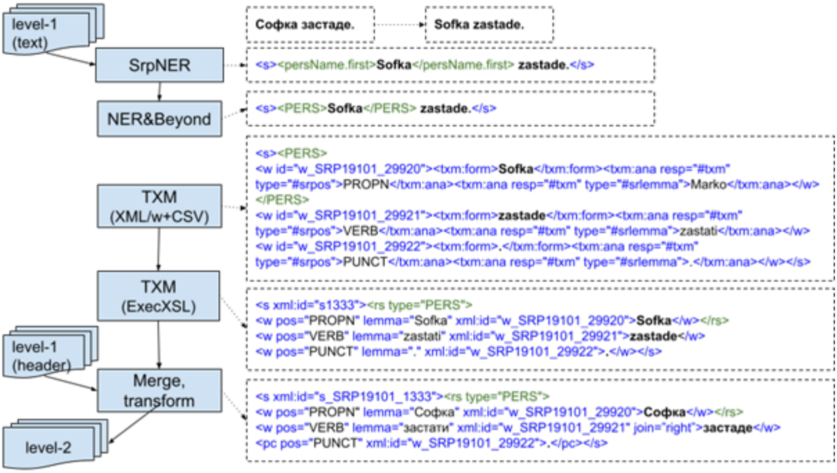
3 The Serbian Level-2 Pipeline

The Serbian level-2 novels are produced from the level-1 edition, as proposed by the Action plan and similarly to the way it was done for some other languages. Each language has its own pipeline, since the best tools for specific languages are developed within different frameworks. For the majority of languages, the integration of morphosyntactic tagging, lemmatization and named entity annotation was not a trivial task. In this section we present the Serbian language pipeline, which comprises several steps of annotations and transformations, outlined in Figure 1, with an example of a short sentence form the well-known novel *Nečista krv* (Impure blood) (SRP19101) by Borisav Stanković.

The TEI document level-1 has elements <teiHeader> and <text> on the first level, but annotation is performed only on the content of the <text>

[4] WG2 data repository

element. For processing purposes the `<teiheader>` element is removed in this phase, only to be updated and merged with the `<text>` element after all annotations are done.



```
<s><persName.first>Sofka</persName.first> zastade.</s>
```

```
<s><PERS>Sofka</PERS> zastade.</s>
```

```
<s><PERS>
<w id="w_SRP19101_29920"><txm:form>Sofka</txm:form><txm:ana resp="#txm"
type="#srpos">PROPN</txm:ana><txm:ana resp="#txm" type="#srlemma">Marko</txm:ana></w>
</PERS>
<w id="w_SRP19101_29921"><txm:form>zastade</txm:form><txm:ana resp="#txm"
type="#srpos">VERB</txm:ana><txm:ana resp="#txm" type="#srlemma">zastati</txm:ana></w>
<w id="w_SRP19101_29922"><txm:form>.</txm:form><txm:ana resp="#txm"
type="#srpos">PUNCT</txm:ana><txm:ana resp="#txm" type="#srlemma">.</txm:ana></w></s>
```

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Sofka" xml:id="w_SRP19101_29920">Sofka</w></rs>
<w pos="VERB" lemma="zastati" xml:id="w_SRP19101_29921">zastade</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.</w></s>
```

```
<s xml:id="s_SRP19101_1333"><rs type="PERS">
<w pos="PROPN" lemma="Софка" xml:id="w_SRP19101_29920">Софка</w></rs>
<w pos="VERB" lemma="застати" xml:id="w_SRP19101_29921" join="right">застаде</w>
<pc pos="PUNCT" xml:id="w_SRP19101_29922">.</pc></s>
```

Figure 1. The Serbian SrpELTeC Level-2 pipeline.

Sentence splitting is performed by a Unitex transducer (Krstev 2008; Paumier 2021) that is adapted for this purpose, to take in consideration tags introduced in level-1. This transducer outputs the start tag `<s>` at the beginning of a sentence and the end tag `</s>` at its end.

The next step is named entity recognition performed by the rule-based system SrpNER (Krstev et al. 2014), based on large-scale lexical resources for Serbian (Krstev 2008), coupled with local grammars in the form of finite-state transducers (Vitas and Krstev 2012). Since SrpNER works on Latin texts, it is necessary to transliterate Cyrillic texts to Latin. SrpNER recognises 11 classes of NEs: dates and time (moments and periods), money and measurement expressions, geopolitical names (countries, settlements, oronyms and hydronyms), and personal names (one or more last names with or without first names and nicknames, names of church and state dignitaries). Here are some examples of SrpNER output:

1. `<pers.spec><role>`carica`</role> <persName.full>`Marija Terezija`</persName.full> </pers.spec>`
   Empress Maria Theresa

2. <pers><role>Sekretar</role> <persName.last>Živanović</persName.last></pers>
   Secretary Živanović
3. <org>Saborna crkva u <top.gr>Beogradu</top.gr></org>
   Cathedral in Belgrade
4. <org>manastir <pers.spec>Sv. Marka</pers.spec></org>
   St. Marc's monastery
5. <pers.spec><role>veliki vezir</role> <persName.first>Ahmed</persName.first>-<role>paša</role></pers.spec>
   Grand Vizier Ahmed-Pasha

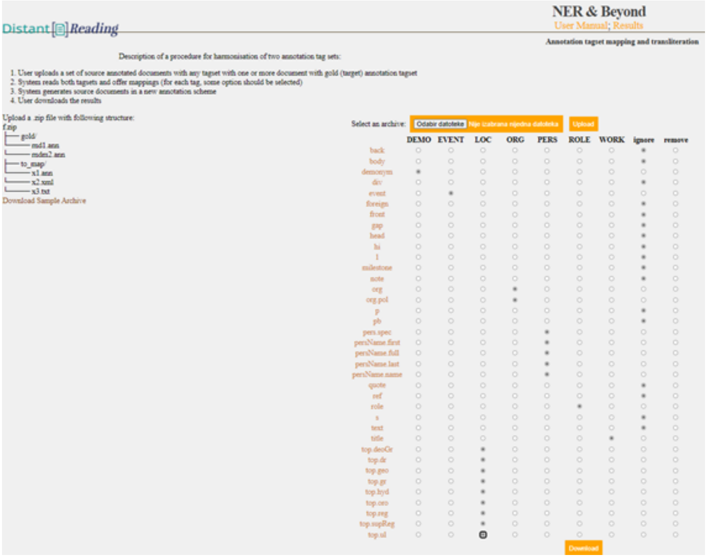
Since level-2 does not allow embedded NER tags, the first step was to apply a semi-automatic procedure to remove them from the SrpNER output. Previous examples would be transformed to:

1. <role>carica</role> <persName.full>Marija Terezija</persName.full>
2. <role>Sekretar</role> <persName.last>Živanović</persName.last>
3. <org>Saborna crkva u Beogradu</org>
4. <org>manastir Sv. Marka</org>
5. <role>veliki vezir</role> <persName.first>Ahmed</persName.first><role>paša</role>

As it can be seen, besides the removal of embedded tags, the remaining SrpNER tags have to be mapped into a more simplified level-2 tagset: PERS, ROLE, LOC, ORG, DEMO, EVENT, WORK. An automatic procedure implemented as part of the NER&Beyond portal (Stanković et al. 2019; Šandrih Todorović et al. 2021) was developed and used to map SrpNER tags into the 7-categories ELTeC NER schema. Figure 2 presents the part of the NER&Beyond portal used for tagsets mapping.

The mapping procedure allows mapping, ignore or removal of XML elements. In this case, the following XML elements are ignored: `<back>`, `<body>`, `<div>`, `<foreign>`, `<front>`, `<gap>`, `<head>`, `<hi>`, `<l>`, `<milestone>`, `<note>`, `<p>`, `<pb>`, `<quote>`, `<ref>`, `<s>`, `<text>`, while the mapping is defined as follows:

– `<persName.first>`, `<persName.full>`, `<persName.last>`, `<persName.name>`, `<pers.spec>` → PERS
– `<top.deoGr>`, `<top.dr>`, `<top.geo>`, `<top.gr>`, `<top.hyd>`, `<top.oro>`, `<top.reg>`, `<top.supReg>`, `<top.ul>` → LOC



Annotation tagset mapping and transliteration

Description of a procedure for harmonisation of two annotation tag sets:

1. User uploads a set of source annotated documents with any tagset with one or more document with gold (target) annotation tagset
2. System reads both tagsets and offer mappings (for each tag, some option should be selected)
3. System generates source documents in a new annotation scheme
4. User downloads the results

Upload a .zip file with following structure:

```
f.zip
├──gold/
│  └──md1.ann
│  └──mden2.ann
├──to_map/
   └──x1.ann
   └──x2.xml
   └──x3.txt
```

Select an archive: Odabir datoteke Nije izabrana nijedna datoteka

	DEMO	EVENT	LOC	ORG	PERS	ROLE	WORK	ignore	remove
back	○	○	○	○	○	○	○	●	○
body	○	○	○	○	○	○	○	●	○
demonym	●	○	○	○	○	○	○	○	○
div	○	○	○	○	○	○	○	●	○
event	○	●	○	○	○	○	○	○	○
foreign	○	○	○	○	○	○	○	●	○
front	○	○	○	○	○	○	○	●	○
gap	○	○	○	○	○	○	○	●	○
head	○	○	○	○	○	○	○	●	○
hi	○	○	○	○	○	○	○	●	○
l	○	○	○	○	○	○	○	●	○
milestone	○	○	○	○	○	○	○	●	○
note	○	○	○	○	○	○	○	●	○
org	○	○	○	●	○	○	○	○	○
org.pol	○	○	○	●	○	○	○	○	○
p	○	○	○	○	○	○	○	●	○
pb	○	○	○	○	○	○	○	●	○
pers.spec	○	○	○	○	●	○	○	○	○
persName.first	○	○	○	○	●	○	○	○	○
persName.full	○	○	○	○	●	○	○	○	○
persName.last	○	○	○	○	●	○	○	○	○
persName.name	○	○	○	○	●	○	○	○	○
quote	○	○	○	○	○	○	○	●	○
ref	○	○	○	○	○	○	○	●	○
role	○	○	○	○	○	●	○	○	○
s	○	○	○	○	○	○	○	●	○
text	○	○	○	○	○	○	○	●	○
title	○	○	○	○	○	○	●	○	○
top.deoGr	○	○	●	○	○	○	○	○	○
top.dr	○	○	●	○	○	○	○	○	○
top.geo	○	○	●	○	○	○	○	○	○
top.gr	○	○	●	○	○	○	○	○	○
top.hyd	○	○	●	○	○	○	○	○	○
top.oro	○	○	●	○	○	○	○	○	○
top.reg	○	○	●	○	○	○	○	○	○
top.supReg	○	○	●	○	○	○	○	○	○
top.ul	○	○	●	○	○	○	○	○	○

Figure 2. The tagsets mapping in the NER&Beyond portal.

- `<demonym>` → DEMO
- `<event>` → EVENT
- `<org>`, `<org.pol>` → ORG
- `<role>` → ROLE
- `<title>` → WORK.

The previous examples would be mapped as follows:

1. `<ROLE>`carica`</ROLE>` `<PERS>`Marija Terezija`</PERS>`
2. `<ROLE>`Sekretar`</ROLE>` `<PERS>`Živanović`</PERS>`
3. `<ORG>`Saborna crkva u Beogradu`</ORG>`
4. `<ORG>`manastir Sv. Marka`</ORG>`
5. `<ROLE>`veliki vezir`</ROLE>` `<PERS>`Ahmed`</PERS>`-`<ROLE>`paša`</ROLE>`

The next step was the preparation of a CSV file with metadata for 100 novels to be used for the import of the whole collection in the TXM tool (Heiden 2010).[5] The TXM import option “XML/w+CSV” was used and the required data supplied: a path to the text collection and metadata, as well as language selection. Namely, depending on language selection, TXM is

[5] Textométrie//TXM

using an appropriate parameter file for TreeTagger, which is used for the part of speech tagging and lemmatization. Tokenization was applied by a set of rules. The Treetagger model[6] was trained using a dataset[7] created from several merged annotated Serbian texts, with over half a million tagged tokens (Table 1). The dataset was balanced with four literary (1-5), and three non-literary (6-8) texts, the former including one complete SrpELTeC novel (3) and a set of excerpts from SrpELTeC (5). Tokens were pre-tagged for Universal POS tagset[8] and lemma with the Unitex system,[9] using Serbian morphological dictionaries, and disambiguated manually. TreeTagger also requires a lexicon and a list of open classes for the training procedure. Serbian morphological dictionaries were used as a lexicon[10] for training, while a list of open classes was used as suggested by the Universal dependencies.

The selection of 11 full novels and excerpts from 15 novels from SrpELTeC, have been automatically labelled with SrpNER system for Serbian in the first stage of the gold standard preparation. Based on the specifically tailored guidelines, different evaluators performed careful checks and corrections, yielding a gold standard (SrpELTeC-gold) that is publicly available on European Language Grid (ELG) platform[11]. Corpus is annotated with 7 different named entity types: PERS, ROLE, LOC, DEMO, ORG, WORK, EVENT, as specified by Distant Reading for European Literary History (COST Action CA16204). Total number of text files is 242 with stend-off annotation in 242 .ann files. Total number of annotations is 330119, where PERS has 14788, ROLE has 10405, LOC has 1979, DEMO 1568, ORG 323, WORK 198, EVENT 149.

A Named Entity Recognizer (SrpCNNER) is trained using SrpELTeC-gold to recognize 7 previously mentioned named entity types, with a Convolutional Neural Network (CNN) architecture, having F1 score of approx 91% on the test dataset. Model trained for spaCy is publicly available on ELG[12].

The benefit of using TXM for tagging is that it retains XML structure elements and adds new information to each token. For example, the sentence *Sofka zastade.* (Sofka paused.):

```
<s><PERS>Sofka</PERS> zastade.</s>
```

[6] SrpKor4Tagging-TreeTagger

[7] SrpKor4Tagging

[8] Universal POS tags

[9] Unitex/GramLab Grammar-based Corpus Processing Suite

[10] SrpMD4Tagging

[11] SrpELTeC-gold - Named Entity Recognition Training corpus for Serbian

[12] SrpCNNER - Named Entity Recognizer for Serbian

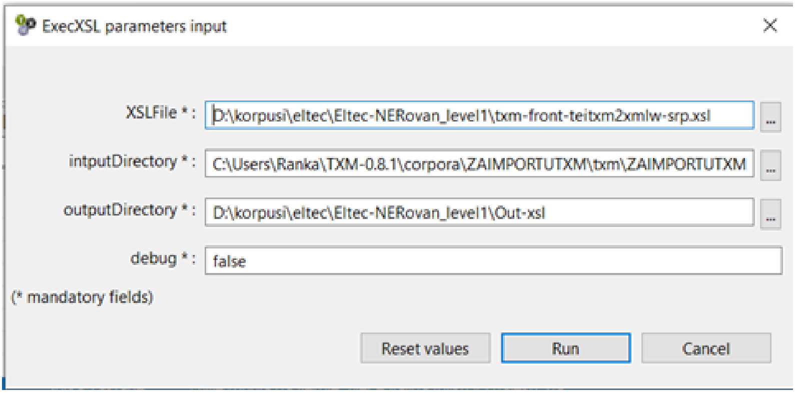
Id	Texts	Tokens	Words	Unique
1	Orwell's *1984* (Serbian translation)	108,137	96,026	18,050
2	Vern's *Around the World in Eighty Days* (Serbian translation)	68,697	62,769	12,799
3	Dragutin Ilić's *Hadži Đera* (SRP19040)	65,262	61,217	12,276
4	Excerpt from Jaroslav Hašek's *The Good Soldier Švejk*	4,122	3,347	1,475
5	Excerpts from *SrpELTeC (1840-1920)*	5,118	4,236	2,093
6	Corpus of newspaper articles on 2014 floods in Serbia	4,672	3,813	1,741
7	Excerpts from the Serbian history textbook	6,596	5,287	2,622
8	A collection of Serbian texts from Law, Finance, Education and Health domain	239,614	204,643	31,470
	Total	502,213	441,338	

Table 1. Annotated texts used for TreeTagger training, as well as the number of tokens, words and unique words for each of them.

becomes:

```
<s><PERS>
  <w id="w_SRP19101_29920"><txm:form>Sofka</txm:form>
    <txm:ana resp="#txm" type="#srpos">PROPN</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">Sofka</txm:ana>
  </w></PERS>
  <w id="w_SRP19101_29921"><txm:form>zastade</txm:form>
    <txm:ana resp="#txm" type="#srpos">VERB</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">zastati</txm:ana>
  </w>
  <w id="w_SRP19101_29922"><txm:form>.</txm:form>
    <txm:ana resp="#txm" type="#srpos">PUNCT</txm:ana>
    <txm:ana resp="#txm" type="#srlemma">.</txm:ana></w>
</s>
```

The obtained result is not yet level-2 compliant, which means that some additional transformations are necessary. Within the TXM tool there is an execXSL macro (in the View→Macro menu within xml macros), which performs transformations. It requires the path to the XSL file, and input and output directory with corpus files that need to be transformed (Figure 3). The initial, general purpose macro *txm-front-teitxm2xmlw.xsl* had to be adapted for the level-2 requirements, and this new version is published on github repository.[13]



XSLFile * : D:\korpusi\eltec\Eltec-NERovan_level1\txm-front-teitxm2xmlw-srp.xsl

intputDirectory * : C:\Users\Ranka\TXM-0.8.1\corpora\ZAIMPORTUTXM\txm\ZAIMPORTUTXM

outputDirectory * : D:\korpusi\eltec\Eltec-NERovan_level1\Out-xsl

debug * : false

(* mandatory fields)

Figure 3. Invocation of the TXM macro for XSL transformation.

The adaptation of the XSL transformation macro included sentence counting, the use of required namespaces for the attributes `xml:id`, `xml:lang`, removing some attributes, and mapping XML elements for NER tags – `<PERS>`, `<LOC>`, `<ORG>`, `<DEMO>`, `<ROLE>`, `<WORK>`, `<EVENT>` – into the referring string TEI element `<rs>`, with the value of its attribute `@type` set to the appropriate value from the set {PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT}. The part of this XSL code is:

```
<xsl:template match= "tei:PERS|tei:LOC|tei:ORG|tei:DEMO|
   tei:ROLE|tei:WORK|tei:EVENT">
   <!-- produce a referring string element -->
   <xsl:element name="rs"
      namespace="http://www.tei-c.org/ns/1.0">
      <xsl:attribute name="type">
         <xsl:value-of select="local-name()"/>
      </xsl:attribute>
      <xsl:apply-templates select="tei:w"/>
      <xsl:apply-templates select="tei:foreign"/>
```

[13] TXM related scripts

```
   </xsl:element>
</xsl:template>
```

As a result, for our example sentence the following would be obtained:

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Sofka" xml:id="w_SRP19101_29920">
  Sofka</w></rs>
<w pos="VERB" lemma="zastati" xml:id="w_SRP19101_29921">
  zastade</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.
</w></s>
```

At the end, some last transformations had to be done. First, the text has to be transformed back to Cyrillic, if that was the script used in level-1, taking care about the content of the `<foreign>` element, which has to be treated in a special way. Since values of all `xml:id` attributes have to be unique for the whole ELTeC collection, the ID of a novel (value of the `xml:id` attribute of the <text> element) needs to be integrated into the sentence ID. Since TEI uses the `<pc>` element, rather than `<w>`, for punctuation, special characters `<w>` elements had to be replaced with `<pc>` and the lemma attribute removed. After this final transformation, our example sentence in the correct level-2 form is:

```
<s xml:id="s1333"><rs type="PERS">
<w pos="PROPN" lemma="Софка" xml:id="w_SRP19101_29920">
  Софка</w></rs>
<w pos="VERB" lemma="застати" xml:id="w_SRP19101_29921">
  застаде</w>
<w pos="PUNCT" lemma="." xml:id="w_SRP19101_29922">.
</w></s>
```

4 Statistical overview of level-2

SrpELTeC level-2 corpus has 100 novels annotated with part of speech tags and lemmas, while 65 novels have also named entity annotation. SrpELTeC has 5,886,528 tokens according to TXM calculation, with the four word properties: word, n, srpos, srlemma and 30 XML tags for structural elements (`back`, `body`, `front`, `div`, `div1`, `div2`, `gap`, `head`, `l`, `milestone`, `note`, `p`, `pb`, `quote`, `ref`, `s`, `text`, `title`, `trailer`), for NER elements (PERS, LOC, ORG, DEMO, ROLE, WORK, EVENT) and other textual elements (`foreign`, `hi`).

Element `<div>` occurs at three levels. At the first level it occurs 1,763 times with the following values of the attribute `@type`: CHAPTER, GROUP, LIMINAL, NOTES, TITLEPAGE. At the second level, it occurs 463 times with chapter or group as values of the attribute @type, while at the third level it occurs 99 times as a chapter. The number of occurrences of other elements are represented in Figure 4. The distribution of named entities is represented in Figure 5.

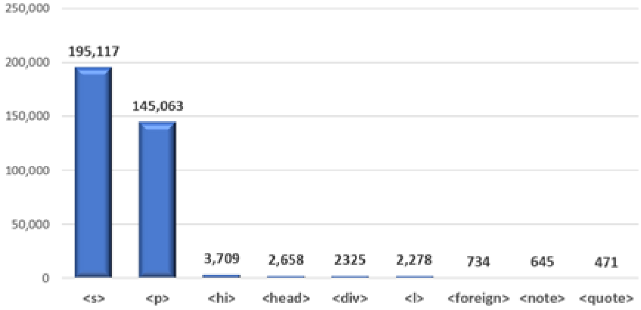


Figure 4. The number of occurrences of elements other than `<div>`.

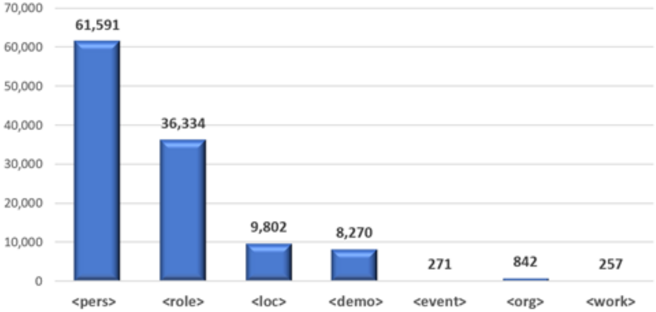


Figure 5. The number of occurrences of NE elements by class.

The average number of words per paragraph is 40, while the average number of words per sentence is 14 (Figure 6). The novels with the longest average sentences are: *Zločin jedne svekrve* (The crime of one mother in law) (SRP19062) (26), *Bespuće* (Wasteland) (SRP19121) (25), *Gmundensko jezero* (Gmunden Lake) (SRP18690) (22). The shortest sentences were used in novels: *Hajduk Stanko* (Haiduk Stanko) (SRP18963) (7), *Radetića Mara* (Radetić's Mara) (SRP18940) (8), *Srbin i Hrvatica* (A Serb and a Croat woman) (SRP18921) (9), *Seljanka* (A peasant woman) (SRP18932) (9). It is interesting to note that *Hajduk Stanko* and *Seljanka* were written by the same author.

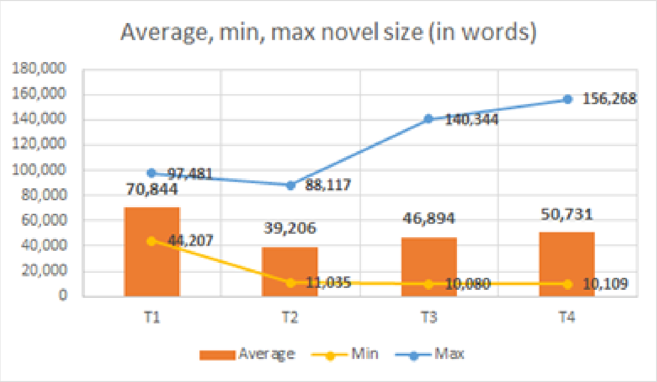


Figure 6. The average size, shortest and longest novels counted by the number of words per time period.

The frequencies for part of speech tags are given in Figure 7, which shows that nouns are the most frequent, followed by verbs and other parts of speech.

The lexicon statistics retrieved by TXM gives insight into most frequently used nouns, verbs, adjectives and pronouns. The 12 most frequent words from each of these groups are given in Table 2. One can see that the most frequent nouns are *ruka* (hand), *kuća* (house) and *dan* (day); the most frequent verbs (apart from auxiliaries) are *moći* (to can), *reći* (to say) and *znati* (to know); the most frequent adjectives are *drugi* (other or second), *velik* (big) and *star* (old); the most frequent pronouns are personal pronouns *on* (he), *ja* (I) and *ona* (she).

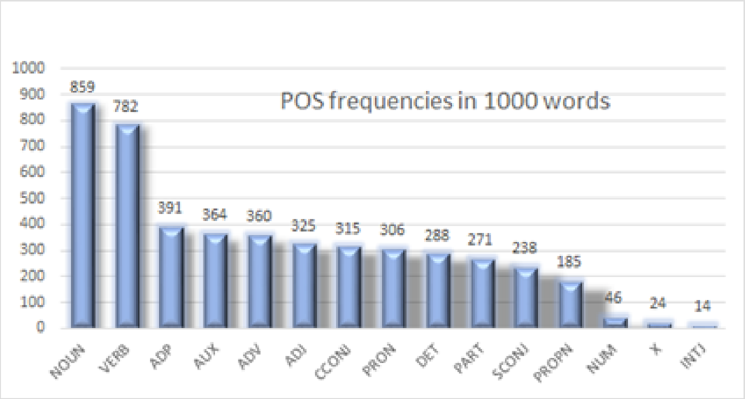


Figure 7. The frequencies of POS in SrpELTeC.

For each novels are calculated absolute frequency (F_i), where (i) represents specific novel and normalized length (Len_i) as the integer division of number of words in novel and 10000, so Len_i are numbers of values that fall in the interval $[1, 15]$. Figure 8 illustrates the most frequent named entities for four categories, using their lemmatized forms. Three frequency values are given for each category: absolute frequency in the whole corpus (green) (F_a), the number of novels in which a NE occurs (divided by 100) (blue) (F_n) and the relative frequency, taking into account both the length of a novel and the number of novels in which a particular NE occurs (yellow) (F_r) calculated using Equation (1).

$$F_r = F_n \cdot \sum_{i=1}^{65} \frac{F_i}{Len_i} \tag{1}$$

The most frequent PERS named entity, both measured by absolute and relative frequency is *Bog* (God). Apart from it the highest absolute frequency have the masculine personal name *Miloš* and feminine name *Darinka*. Measured by number of novels in which they occur, the most used are the masculine personal name *Pera* and feminine name *Mara*. The most frequent ROLE entities are *gospodin* (mister), *pop* (priest) and *gospođa* (missis), measured both by absolute and relative frequency. The roles that appear in most of the novels, besides *gospodin*, are *seljak* (peasant) and *gazda* (landlord). Other frequent roles are *kapetan* (captain), *učitelj* (teacher) and *kmet* (farmer).

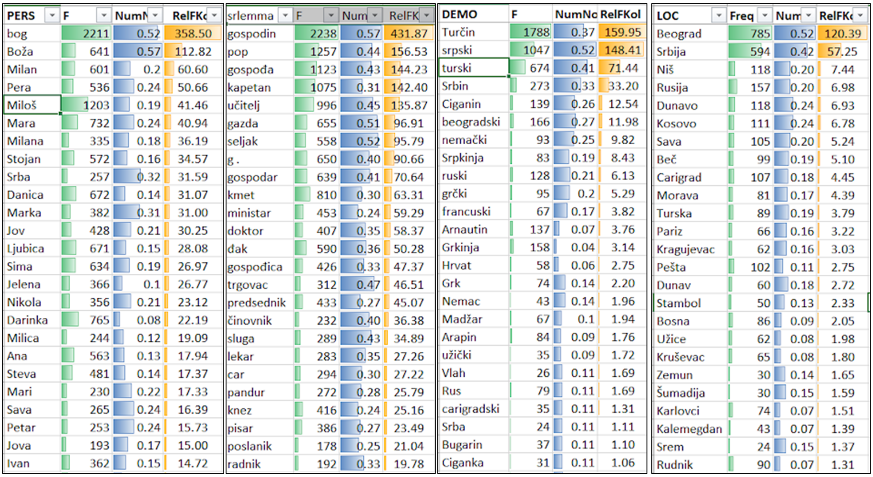
As for DEMO entities, the entities referring to inhabitants or ethnic groups having the highest relative frequency are *Turci* (Turks), *Srbi* (Serbians) and

NOUN	FREQ	VERB	FREQ	ADJ	FREQ	PRON	FREQ
ruka	10,802	moći	20,353	drugi	12,553	on	89,345
kuća	10,461	reći	18,215	velik	8,341	ja	60,830
dan	9,238	znati	15,399	star	5197	ona	56,925
glava	8,046	imati	13,432	dobar	4,774	ti	23,599
oči	7,841	doći	9,106	prvi	4,773	vi	13,707
bog	6,998	videti	8,252	mlad	4,474	šta	12,547
put	6,370	ići	7247	ceo	4,314	mi	8,299
čovek	6,019	kazati	7,065	lep	4,151	ko	8,274
ljudi	5,586	početi	6,780	nov	3,706	sebe	8058
reč	5,141	govoriti	6,462	crn	2,961	oni	5,262
žena	4,923	gledati	6,410	srpski	2,908	ništa	3,036
vreme	4,825	misliti	6,206	mali	2,887	ono	2,764

Table 2. The most frequent nouns, verbs, adjectives and pronouns

Cigani (Roma people). The most freqeunt adjectives referring to toponyms, inhabitants or ethnic groups are *srpski* (referring to Serbia or Serbians), *turski* (referring to Turkey or Turks) and *beogradski* (referring to Belgrade). The most frequent LOC entities both measured by absolute and relative frequency are *Beograd* (Belgrade) and *Srbija* (Serbia). Besides them, the frequently occurring countries are *Rusija* (Russia) and *Turska* (Turkey), the frequently occurring cities in Serbia are *Niš*, *Kragujevac* and *Užice*, the cities that are not in Serbia are *Beč* (Vienna), *Carigrad/Stambol* (Istanbul) and *Pariz* (Paris), and the most frequent rivers are *Dunav* (Danube), *Sava* and *Morava*.

Multi-word units are not annotated in the level-2 version of ELTeC collection, except for the named entities. Due to the existence of the incomplete morphological dictionaries of multi-word units we were able to retrieve the most frequently used multi-word nouns and adjectives. By far the most frequent multi-word noun is *srpski narod* (Serbian people), followed by *bojno polje* (battle field) and *vrhovna komanda* (High Command). The frequent multi-word nouns referring to education are *osnovna škola* (elementary school), *školska godina* (school year) and *učitelj muzike* (music teacher). It is interesting that adjectives *železnički* (referring to railway) and *električni* (electric) are used in numerous multi-word nouns revealing the modernization of Serbia: *železnička pruga* (railway) and *železnička stan-*



PERS	F	NumN	RelFKc
bog	2211	0.52	358.50
Boža	641	0.57	112.82
Milan	601	0.2	60.60
Pera	536	0.24	50.66
Miloš	1203	0.19	41.46
Mara	732	0.24	40.94
Milana	335	0.18	36.19
Stojan	572	0.16	34.57
Srba	257	0.32	31.59
Danica	672	0.14	31.07
Marka	382	0.31	31.00
Jov	428	0.21	30.25
Ljubica	671	0.15	28.08
Sima	634	0.19	26.97
Jelena	366	0.1	26.77
Nikola	356	0.21	23.12
Darinka	765	0.08	22.19
Milica	244	0.12	19.09
Ana	563	0.13	17.94
Steva	481	0.14	17.37
Mari	230	0.22	17.33
Sava	265	0.24	16.39
Petar	253	0.24	15.73
Jova	193	0.17	15.00
Ivan	362	0.15	14.72

srlemma	F	Num	RelFK
gospodin	2238	0.57	431.87
pop	1257	0.44	156.53
gospođa	1123	0.43	144.23
kapetan	1075	0.31	142.40
učitelj	996	0.45	135.87
gazda	655	0.51	96.91
seljak	558	0.52	95.79
g .	650	0.40	90.66
gospodar	639	0.41	70.64
kmet	810	0.30	63.31
ministar	453	0.24	59.29
doktor	407	0.35	58.37
đak	590	0.36	50.28
gospođica	426	0.33	47.37
trgovac	312	0.47	46.51
predsednik	433	0.27	45.07
činovnik	232	0.40	36.38
sluga	289	0.43	34.89
lekar	283	0.35	27.26
car	294	0.30	27.22
pandur	272	0.28	25.79
knez	416	0.24	25.16
pisar	386	0.27	23.49
poslanik	178	0.25	21.04
radnik	192	0.33	19.78

DEMO	F	NumNo	RelFKol
Turčin	1788	0.37	159.95
srpski	1047	0.52	148.41
turski	674	0.41	71.44
Srbin	273	0.33	33.20
Ciganin	139	0.26	12.54
beogradski	166	0.27	11.98
nemački	93	0.25	9.82
Srpkinja	83	0.19	8.43
ruski	128	0.21	6.13
grčki	95	0.2	5.29
francuski	67	0.17	3.82
Arnautin	137	0.07	3.76
Grkinja	158	0.04	3.14
Hrvat	58	0.06	2.75
Grk	74	0.14	2.20
Nemac	43	0.14	1.96
Madžar	67	0.1	1.94
Arapin	84	0.09	1.76
užički	35	0.09	1.72
Vlah	26	0.11	1.69
Rus	79	0.11	1.69
carigradski	35	0.11	1.31
Srba	24	0.11	1.11
Bugarin	37	0.11	1.10
Ciganka	31	0.11	1.06

LOC	Freq	Num	RelFKc
Beograd	785	0.52	120.39
Srbija	594	0.42	57.25
Niš	118	0.20	7.44
Rusija	157	0.20	6.98
Dunavo	118	0.24	6.93
Kosovo	111	0.24	6.78
Sava	105	0.20	5.24
Beč	99	0.19	5.10
Carigrad	107	0.18	4.45
Morava	81	0.17	4.39
Turska	89	0.19	3.79
Pariz	66	0.16	3.22
Kragujevac	62	0.16	3.03
Pešta	102	0.11	2.75
Dunav	60	0.18	2.72
Stambol	50	0.13	2.33
Bosna	86	0.09	2.05
Užice	62	0.08	1.98
Kruševac	65	0.08	1.80
Zemun	30	0.14	1.65
Šumadija	30	0.15	1.59
Karlovci	74	0.07	1.51
Kalemegdan	43	0.07	1.39
Srem	24	0.15	1.37
Rudnik	90	0.07	1.31

Figure 8. The frequencies of PERS, ROLE, DEMO and LOC categories in 65 novels of SrpELTeC.

ica (railway station), *električna struja* (electric current), *električna baterija* (electric battery), *električna lampa* (electric lamp), *električna sijalica* (electric bulb), *električno zvonce* (electric bell) and *električna centrala* (electric power station). Frequently occurring multi-word nouns with figurative meaning are *mrtva tišina*, (dead silence) *grobna tišina* (grave silence) and *crne misli* (black thoughts). Among multi-word adjectives, excluding similes (see (Krstev 2021) in the same issue) and demonyms are: *živ i zdrav* (alive and healthy), *go i bos* (nude and barefoot), *mrtav pijan* (deadly drunk), *mrtav umoran* (deadly tired).

5 Conclusions

In this paper we presented the results of the team work on producing the so-called level-2 edition of SrpELTeC. We gave an overview of the required schema with its main characteristics, and challenges in processing. Serbian level-2 pipeline included adaptation of SrpNER for named entity annotation, preparation of TreeTagger model for Serbian with the Universal Dependencies tagset, part of speech annotation and lemmatization within TXM tool, and preparation of several scripts for file transformations. Finally, statistics generated by TXM are supplied for several tags used as structural elements. Statistics are generated by using TXM.

Further plans include NER annotation of remaining 35 novels of SrpELTeC and adaptation of the output format to be compliant with the final level-2 schema, which is expected soon. The addition of the new layer with multi-word units annotation is also envisaged. The srpELTeC corpus will be further analysed by the quantitative and qualitative approach to researching textual corpus elements within the TXM program with the textometric approach (Heiden 2010; Jaćimović 2019) and visual presentation of the obtained results, as well as Latent semantic analysis. Various other analyses will be possible with this valuable resource, like authorship attribution, the lexical attraction between words (co-occurrence analysis), text specificity analysis, MWE and collocation extraction, dictionary example extractions, named entity linking, sentiment analysis, direct speech, word embeddings.

Acknowledgment

We would like to thank Prof. Serge Heiden and his team from the IHRIM, École normale supérieure de Lyon for helping us adapt TXM platform for the purpose of this work. We are also grateful to Prof. Diana Santos from the Department of Literature, Area Studies and European Languages at University of Oslo for many fruitful discussions that helped in finding best solutions to various problems.

References

Burnard, Lou, Christof Schöch, and Carolin Odebrecht. 2021. “In search of comity: TEI for distant reading.” *Journal of the Text Encoding Initiative,* no. 14.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. “Named entity recognition for distant reading in ELTeC.” In *CLARIN Annual Conference 2020.*

Heiden, Serge. 2010. “The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme.” In *24th Pacific Asia conference on language, information and computation,* 2:389–398. 3. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Heiden, Serge, Jean-Philippe Magué, and Bénédicte Pincemin. 2010. "TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement." In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010,* 2:1021–1032. 3. Edizioni Universitarie di Lettere Economia Diritto.

Jaćimović, Jelena. 2019. "Textometric methods and the TXM platform for corpus analysis and visual presentation." *Infotheca* 19 (1): 30–54. https://doi.org/10.18485/infotheca.2019.19.1.2.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana. 2021. "White as Snow, Black as Night – Similes in Old Serbian Literary Texts." *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.6.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489.

Paumier, Sebastian. 2021. *Unitex 3.3 User Manual.* Université Paris-Est Marne-la-Vallée. https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER& Beyond: The Archaic and the Modern Intertwined." In *Deep Learning Natural Language Processing Methods and Applications – Proc. of the Int. Conf. Recent Advances in Natural Language Processing (RANLP 2021),* edited by Galia et al. Angelova, 1252–1260. INCOMA Ltd. https://doi.org/10.26615/978-954-452-072-4_141.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. "Named Entity Recognition for Distant Reading in Several European Literatures." *DH Budapest 2019.*

Vitas, Duško, and Cvetana Krstev. 2012. "Processing of corpora of serbian using electronic dictionaries." *Prace Filologiczne* 63:279–292.

Serbian ELTeC Sub-collection in Wikidata

UDC 004.62: [030:004.738.5
DOI 10.18485/infotheca.2021.21.2.4

Milica Ikonić Nešić
milica.ikonic.nesic@fil.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs

Biljana Rujević
biljana.rujevic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ABSTRACT: This paper presents an example of integration of Wikidata with digital libraries and external systems, as well as some best practices for speeding up the process of data preparation and import to Wikidata, on the use case of SrpELTeC, Serbian subcollection of the ELTeC multilingual collection (European Literary Text Collection). After preliminary work on the manual Wikidata population with SrpELTeC novels, the goal was to automate the process of preparing and importing information, so different solutions were analysed and finally synergy of two, OpenRefine and QuickStatements, was chosen as the best option. The paper also brings examples of SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualisation options.

KEYWORDS: Wikidata, distant reading, literary corpus, named entity linking, ELTeC, SrpELTeC.

PAPER SUBMITTED: 05 November 2021
PAPER ACCEPTED: 12 November 2021

1 Introduction

Wikidata[1] is a Wikimedia Foundation knowledge base, a common source of various kinds of data, used not only by other Wikimedia projects, but also increasingly by numerous semantic web applications. Integration of Wikidata with digital libraries and external systems is envisaged as a useful task for various applications. Wikidata has grown significantly since its launch in October 2012. It has also become the most edited Wikimedia project,

[1] Wikidata

supporting 150–500 edits per minute, or half a million per day— about three times as many as the English Wikipedia. About 90% of these edits are made by bots created by contributors to automate tasks, yet almost one million edits per month are still made by humans (Vrandečić and Krötzsch 2014). It supports over 350 languages, especially English, Dutch, French and German, contains more than 200 million statements on about 56 million items and has a higher edit frequency than Wikipedia,[2] and in this work it is used for Serbian.

Wikidata, as an open data network, was used by Andonovski (Андоновски 2019) to describe language resources, namely, novels from the Serbian-German literary corpus (Andonovski, Šandrih, and Kitanović 2019). Stanković and Davidović (2021) presented an example of integration of Wikidata with digital libraries and external systems, as well as the potential for speeding up the process of data preparation and entry, using articles published in the journal for digital humanities *Infotheca*, as an example. Wikidata's popularity in medicine and bioinformatics is also growing very fast. The potential use of Wikidata as a useful resource for biomedical data integration and semantic interoperability between biomedical computer systems is rising. Different knowledge resources can be automatically processed by users as well as by computer methods and programs, and it was shown how that can be useful for various medical purposes such as clinical decision support (Turki et al. 2019). The Scholia[3] project (Nielsen, Mietchen, and Willighagen 2017) is one of the first comprehensive endeavours of its kind aimed at representing bibliographical data, scholarly profiles of authors and institutions using Wikidata. To the best of our knowledge, this work is the first example of automatically imported data about literary text corpus in Wikidata using different open source tools. The main concepts and usage of Wikidata in our research are presented in Section 2.

The opportunity for speeding up the process of data preparation was seen in using information already encoded in the header of each novel (Krstev 2021) in the SrpELTeC, a subcollection of ELTeC – European Literary Text Collection[4] of novels from the period 1840-1920, developed as part of the "Distant Reading for European Literary History Cost Action"[5] (COST Ac-

[2] Language Statistica for Items

[3] Scholia, Scholia in Wikidata

[4] ELTeC (Distant Reading for European Literary History)

[5] D-reading home page

tion CA16204) by members of the JeRTeh society, led by Cvetana Krstev and Ranka Stanković (Stanković et al. 2019; Frontini et al. 2020).

Cooperation of Wikimedia Serbia[6] and their education program[7] with the University of Belgrade has a long tradition. Work on entering metadata about Serbian novels from the SrpELTeC corpus (Krstev et al. 2019) and linking Wikidata to various applications, one of which is *Aurora*,[8] has been going on for many years. Students of the Faculty of Mining and Geology are trained to populate and use Wikidata, and the application possibilities of open data are studied within the subject Presentation of Knowledge and the Semantic Web at the University of Belgrade multidisciplinary studies PhD program Intelligent Systems. Before the activities described in this paper, entry of ELTeC metadata was manual.

A set of metadata of the SrpELTeC novels, which will be presented in the third section of this paper, is extracted from the `<TEIHeader>` element, to fit the requirements of the ELTeC action schema.[9] Mapping between dataset selected from metadata defined by DR WG1 (Distant Reading working group 1) and Wikidata will be presented, as well as possibilities related to some further, optional data, such as novel's main characters, important places etc. In Section 3 Wikidata concepts will be presented and illustrated by SrpELTeC novel entities and their properties.

Since the automation of the data preparation and import process was envisaged, different solutions were analysed and finally synergy of OpenRefine[10] and QuickStatements[11] tools was chosen as the best option, similar to the approach presented in (Stanković and Davidović 2021). Elaboration of the automation process is given in Section 4.

After completing the SrpELTeC novels Wikidata, a set of web pages integrated results from different queries with different visualisation options, based on Wikidata Query Service, with the Aurora, but further integration with other systems is envisaged. Queries were written that supplied the tables: the title of the novel, the name of the author, the author's pictures,

[6] Wikimedia Serbia

[7] Wikimedia Educational Program

[8] Aurora

[9] ELTeC XML Schemas

[10] OpenRefine, is a tool for working with messy data: cleaning, converting from one format to another, with the addition of external data via a web service, web page

[11] QuickStatements, Wikidata editor: add and remove statements, tags, descriptions, etc., web page

the year of publication, the main characters, the author's distributions by gender, etc. The experience with using SPARQL[12] for data validation will be shared in Section 5. Data on the main characters include elementary data, which should be supplemented with new content in the future: whether the characters are fictional or not, and if they are not, their short biography.

2 Wikidata

Tim Berners Lee believed that the web will enable machines to comprehend semantic documents and data, but not human speech and writings. Properly designed, the semantic web can assist the evolution of human knowledge as a whole (Berners-Lee, Hendler, and Lassila 2001). Nowadays, the semantic web is an extension of the existing web, where information is given a precisely defined meaning, and which enables better cooperation between computers and users. The concept of the semantic web and open related data technologies extend the traditional web using standard markup language and supporting processing tools, where the RDF (Resource Description Framework),[13] a framework for describing resources on the web, plays a major role and provides more efficient solutions for finding information (Shah et al. 2002). For the semantic web operability, computers need to have access to structured collections of information and establish defined rules for automated management. Wikidata is fitting into these trends in information technology development, which are pushing the boundaries from machine readability to machine comprehensibility (understanding) of data on the web, namely from web of documents to web of data. The underlying structure of any expression (statement) in RDF is a collection of triples, each consisting of a subject, a predicate, and an object.

Wikidata is document-oriented, item-centered, representing topics, concepts, or objects and consist of two types of entities: items (e.g. https://www.wikidata.org/wiki/Q107648205) and properties (e.g. https://www.wikidata.org/wiki/Property:P1433). Each item is assigned a unique, permanent identifier "QID" or Q number, which is the unique identifier of a data item on Wikidata, comprising the letter "Q" followed by one or more digits. It is used to help people and machines understand the difference between items with the same or similar names, but different meanings. This number appears next to the name at the top of each Wikidata item. Properties

[12] SPARQL

[13] RDF

cannot be directly created by regular users, to prevent duplication and disorganization of Wikidata properties (e.g. Risk factor property proposal).[14]

The subject of the triple is the Wikidata item to which the claim refers, the predicate is a Wikidata property, and the object is a value. A value can be another item, a string, a time, a period, a location, an URL, or a quantity, depending on the property type. Statements can be made more precise using qualifiers. These qualifiers show the contexts of the validity of the statement. Statements can be annotated by including references. Qualifiers and references are also represented in the form of triples, where the subject is the claim. A claim and its references are considered a statement (Turki et al. 2019). Statements are how any information known about an item is recorded in Wikidata.

The items and properties in Wikidata that are used to structure the ontology are class (Q16889133), entity (Q35120), Wikidata metaclass (Q19361238), instance of (P31), and subclass of (P279). Classes are items that conceptually group together similar items, as human (Q5) groups together humans.

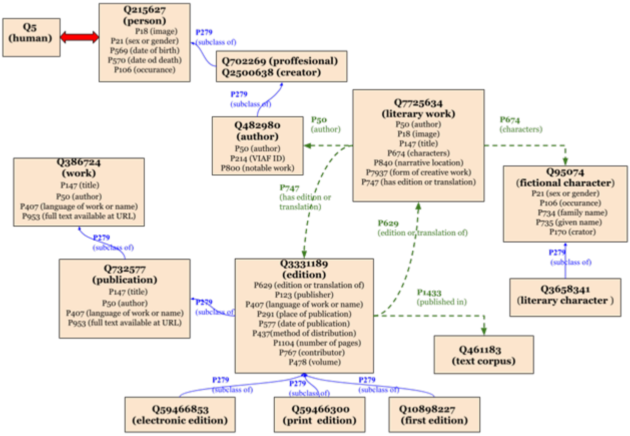


Figure 1. The class diagram of Wikidata used for novels in SrpELTeC.

Figure 1 presents the class/instance relation of all classes and relations that are used in this paper and that will be presented in more detail in

[14] Property proposal/Creative work

Section 3. Blue lines represent "instance of" relations between classes, while green lines present properties that connect two classes by other relations. It is very important to notice that, for the purpose of our work, the class person (Q215627), which represents a "being that has certain capacities or attributes constituting personhood" is used as class humans (Q5) with property "instance of" (P31) as recommended in Wikidata documentation.[15] The class properties are used to describe items in our work, and Figure 1 emphasizes their usage. A very important aspect of the presented work is that each novel is represented with three different items in Wikidata, which will be associated with the appropriate properties, as explained in Subsection 3.2. We will just mention here that some properties like P18 (image), P1104 (number of words), P214 (VIAF[16] ID), occurrence (P106) are general Wikidata properties and P478 (volume) is a qualifier of edition.

3 SrpELTeC Data Structure in Wikidata

To automate the entry of items in Wikidata, the first step was to retrieve and prepare the data, as presented in Subsection 3.1. The second step involved selecting XML elements and attributes in the data to be used to identify predicates and create an input schema. The schema defines the relationship of the value to the item, i.e., the subject, using predicates as intermediaries. Wikidata concepts exemplified by new SrpELTeC entities and their properties are introduced in Section 3.2.

3.1 Data Preparation and Mapping with Wikidata

The work on the ELTeC corpus envisages a basic annotation for all subcollections, the so-called level 1 annotation. The annotation consists of marking the basic structural elements of the text (chapters and other units) and some basic textual elements. The annotation was performed in accordance with the TEI recommendations (TEI Consortium 2021), where from a rich set of elements defined by these recommendations, only a subset was selected as mandatory or allowed.[17]

A metadata header `<TeiHeader>` is required for each text annotated in accordance with TEI recommendations, so it is also the case with all ELTeC

[15] Class person

[16] The Virtual International Authority File – an international reference file for authors and books that includes bibliographic and subject metadata (Loesch 2011).

[17] Encoding Guidelines for the ELTeC: level 1

corpus novels. The mandatory header elements are uniform for all collections and they must contain:

- Description of the electronic edition, which includes the title of the work and the name of the author, as well as the statements of responsibility (scanning, correction, annotation), date of publication, size (measured by the number of words). The author and the work can be joined by identifiers, such as viaf and Wikidata.
- A brief catalog description of the first edition and the edition used as the source for ELTeC (if different from the first edition).
- Description of the text in terms of meeting the balance criteria (e.g. author gender, size, time...) (see (Trtovac, Milnović, and Krstev 2021) from this issue).
- Review of all changes to the digital edition since its first publication.

The first column of Table 1 represents the properties of Wikidata that are used in statements. The XPath[18] expressions used to retrieve the metadata from the TEI header are in the second column. The extracted elements from the TEI header are the values of the property in the statement, where value type can be item, URL or string. All extracted data were labeled, the same properties labeled in the same column. The name in the third column is used in further processing and automation steps. The last column contains information about the class of the instantiated data that is used for mapping, which will be explained later.[19]

Currently exists on Wikidata as a property	**TEI XPath to element (attribute)**	**Name of a column in prepared data**	**Instance of**
	//fileDesc		
	/titleStmt		
P1476 (title)	/title	Title	Q783521 (title)
P214 (viaf id)	/title@ref	_ViafID	Q19832964 (VIAF ID)
P50 (author)	/author	Author	Q482980 (author)
P214 (viaf id)	/author@ref	_ViafID	Q19832964 (VIAF ID)
	/extent		
P657 (number of words)	/measure@unit	Words	Q8034324 (word count)
P1104 (number of pages)	/measure@unit	Pages	Q1069725 (page)
	/publicationStmt		
P123 (publisher)	/publisher	Publisher	Q105044823 (publisher)
P750 (distributed by)	/distributor	Distributor	Q60614978 (distributor)

[18] XML Path Language (XPath) 3.1

[19] Properties table

Currently exists on Wikidata as a property	**TEI XPath to element (attribute)**	**Name of a column in prepared data**	**Instance of**
P6216 (copyright status)	/availability /licence @target	Licence	Q20007257 (CC BY 4.0)
	SourceDesc/bibl	FirstEdition	
P50(author)	/author	_author	Q482980 (author)
P146 (title)	/title	_title	Q783521 (title)
P291 (place of publication)	/pubPlace	_pubPlace	Q1361759 (place of publication)
P123 (publisher)	/publisher	_publisher	Q105044823 (publisher)
P577 (publication date)	/date	_date	Q1361758 (date of publication)
	//profileDesc		
P407 (language of work or name)	langUsage /language	Language	Q34770 (language)
P21 (sex or gender)	/textDesc authorGender@key	authorGender	Q290 (sex)

Table 1. Extraction of information from metadata header.

In the first step of automation, Wikidata items were added for all novels that are in the SrpELTeC collection (more in Section 4), where each novel was created as an instance of literary work (Q7725634), and related with its editions. The editions of the novel, using property P747 (has edition or translation), are connected with a novel with property P629 (edition or translation of). As shown in Figure 2, the data from the first edition and from the ELTeC edition are extracted from the TEI header and mapped to appropriate Wikidata properties, entities and values. The properties that are used to create new items for novels are some of those presented in Table 1, such as P50 (author), P146 (title), P407 (language of work or name) and also new ones such as P674 (characters) and P840 (narrative location), which will be explained later. The data for authors are extracted from TEI header and mapped to Wikidata properties, such as sex or gender (P21), date of birth (P569), date of death (P570), and VIAF ID (P214). One should note that the author's date of birth (P569) and the date of death (P570) are extracted from the author element in the header (framed red in Figure 2).

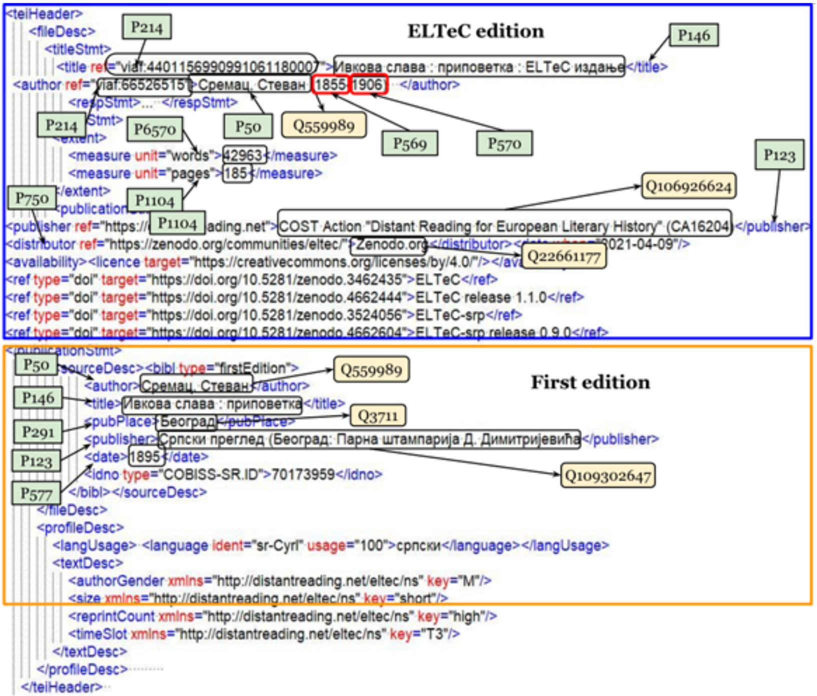
Figure 2 shows an example of mapping between the metadata header of a novel *Ivkova slava* (Ivko's feast) (SRP18950) and Wikidata. Green boxes represent properties that are prefixed by P (e.g. P214 (VIAF ID), P146 (title)) and that are pointing to xml elements or attributes. The content that is framed represent values in the Wikidata statement, and if they have their own QID, they are associated with an appropriate Q identifier (e.g. Stevan Sremac (Q559989), Beograd (Q3711)). The contents that are framed but are not associated with a Q identifier (e.g. 185 (number of words), Ivkova slava: pripovetka: ELTeC izdanje (title), 1895 (date of publication)) are literals. The blue box displays information used to create item ELTeC edition (subclass of edition (Q3331189)) of a novel in Wikidata (e.g. "Ivkova slava: pripovetka: ELTeC izdanje" (Q107648205)). The orange contains information used to create the first edition of a novel in Wikidata (e.g. "Ivkova slava: pripovetka" (Q109336719)).

The narrative characters from a literary work and places where the action takes place can be found in Wikidata for well described novels (e.g. Romeo and Juliet (Q83186); Don Quixote (Q480)). This information is not a part of the metadata header and other extraction methods are required. The SrpELTeC is published in the so-called level-2[20] as well, which supplies more detailed information by annotating all words in the text with their part-of-speech, lemma (word's vocabulary headword form), and optionally other morphosyntactic descriptions, as well as by annotating named entities.

The main goal of named entity recognition in general, is to indicate in a text names of persons, their roles, locations, organizations, and other relevant entities for specific purposes. The first system for recognizing named entities for Serbian is based on manually created rules, which rely on comprehensive Serbian lexical resources (Krstev et al. 2014).

At the level of the whole action, it was agreed that only 7 categories of entities should be indicated in the novels: PERS, ROLE, DEMO, ORG, LOC, WORK, EVENT, which were assessed as being of the greatest importance for further literary studies (Frontini et al. 2020). For the purpose of the work presented here, only two categories PERS and LOC are used. In the list of the extracted PERS entities the main characters of the novel can be found, while in the LOC entity list one expects to find where the narrative of the novel is set. All entities in both categories were sorted by frequency of occurrence in each novel, and the most frequently entities in the PERS category are taken as literary characters (Q3658341), while the most frequently entities in the LOC category are taken as narrative places, i.e. geographic location (Q2221906). This task cannot be fully automated, since names of the same character can be mentioned in a text in a number of different ways, such as: *Ivko*, *Ivka*, *Ivku*, *Ivko Mijalković*.

[20] Encoding Guidelines for the ELTeC: level 2

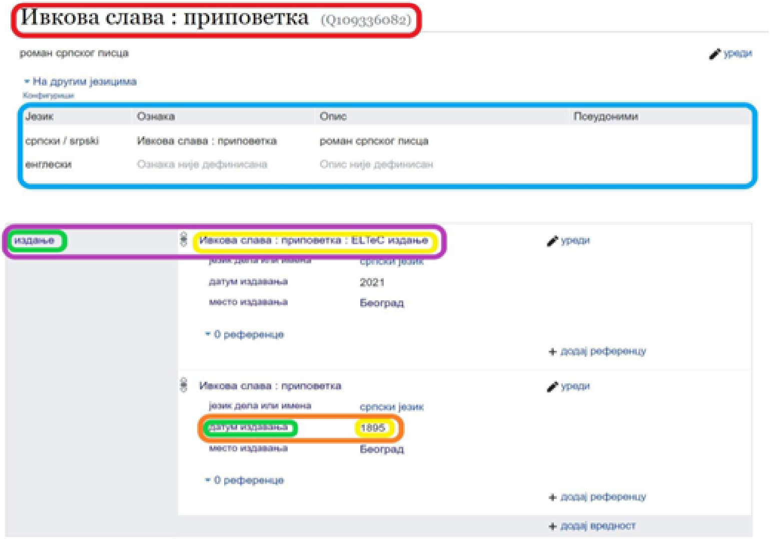


```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title ref="viaf:4401156990991061180007">Ивкова слава : приповетка : ELTeC издање</title>
<author ref="viaf:66526515">Сремац, Стеван (1855-1906)  </author>
      <respStmt>... </respStmt>
    Stmt>
    <extent>
      <measure unit="words">42963</measure>
      <measure unit="pages">185</measure>
    </extent>
    <publication
<publisher ref="https://d ading.net">COST Action "Distant Reading for European Literary History" (CA16204)</publisher>
<distributor ref="https://zenodo.org/communities/eltec/">Zenodo.org</distributor><date when="2021-04-09"/>
<availability><licence target="https://creativecommons.org/licenses/by/4.0/"/></ava
<ref type="doi" target="https://doi.org/10.5281/zenodo.3462435">ELTeC</ref>
<ref type="doi" target="https://doi.org/10.5281/zenodo.4662444">ELTeC release 1.1.0</ref>
<ref type="doi" target="https://doi.org/10.5281/zenodo.3524056">ELTeC-srp</ref>
<ref type="doi" target="https://doi.org/10.5281/zenodo.4662604">ELTeC-srp release 0.9.0</ref>
</publicationStmt>
    <sourceDesc><bibl type="firstEdition">
      <author>Сремац, Стеван</author>
      <title>Ивкова слава : приповетка</title>
      <pubPlace>Београд</pubPlace>
      <publisher>Српски преглед (Београд: Парна штампарија Д. Димитријевића</publisher>
      <date>1895</date>
      <idno type="COBISS-SR.ID">70173959</idno>
    </bibl></sourceDesc>
  </fileDesc>
  <profileDesc>
    <langUsage> <language ident="sr-Cyrl" usage="100">српски</language></langUsage>
    <textDesc>
      <authorGender xmlns="http://distantreading.net/eltec/ns" key="M"/>
      <size xmlns="http://distantreading.net/eltec/ns" key="short"/>
      <reprintCount xmlns="http://distantreading.net/eltec/ns" key="high"/>
      <timeSlot xmlns="http://distantreading.net/eltec/ns" key="T3"/>
    </textDesc>
  </profileDesc>
</teiHeader>
```

Figure 2. Mapping between metadata header and Wikidata.

3.2 Structure of SrpELTeC Wikidata Items

The structure of Wikidata statements allows encoding of the basic information needed to identify the topic covered by an item, without favoring any language, in order to ensure the uniqueness of the meaning of a particular term. Some examples of items used are Beograd (Q3711), Srbija (Q47561), Ivo Andrić (Q47561), Ivkova slava (Q107648205), ELTeC collection (Q106927517). It happens sometimes that there are two items under the same name, e.g. Ivkova Slava (Q107648205), which represents the novel by Stevan Sremac (Q559989), and Ivkova Slava (Q12752161), which represents the movie based on the novel and directed by Zdravko Šotra (Q1253494). It is recommended that in the case of ambiguous entities additional clarification is given in parentheses, such as Ivkova slava (movie). Thus, the item is associated with a unique identifier (QID), while the identifier is associated with a pair: title and description, in order to remove any ambiguity. In our work we used several properties as explained in Subsection 3.1 and some of these properties are used both for novels (literary work) and authors as instance of (P31) and VIAF ID (P214).



Ивкова слава : приповетка (Q109336082)

роман српског писца

Језик	Ознака	Опис	Псеудоними
српски / srpski	Ивкова слава : приповетка	роман српског писца	
енглески	Ознака није дефинисана	Опис није дефинисан	

издање

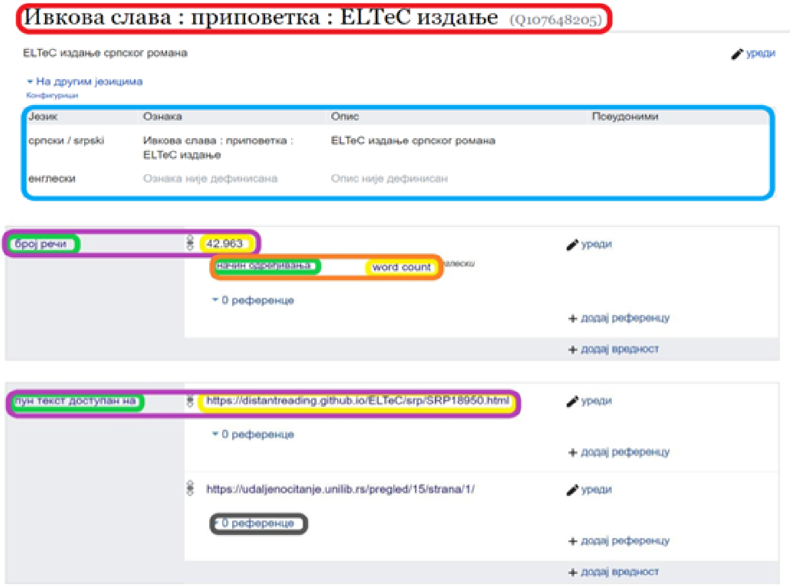
Ивкова слава : приповетка : ELTeC издање

језик дела или имена — српски језик
датум издавања — 2021
место издавања — Београд

Ивкова слава : приповетка

језик дела или имена — српски језик
датум издавања — 1895
место издавања — Београд

Figure 3. Wikidata data structure: the case of a novel.



Ивкова слава : приповетка : ELTeC издање (Q107648205)

ELTeC издање српског романа

Језик	Ознака	Опис	Псеудоними
српски / srpski	Ивкова слава : приповетка : ELTeC издање	ELTeC издање српског романа	
енглески	Ознака није дефинисана	Опис није дефинисан	

број речи: 42.963
начин одређивања: word count

пун текст доступан на: https://distantreading.github.io/ELTeC/srp/SRP18950.html
https://udaljenocitanje.unilib.rs/pregled/15/strana/1/

Figure 4. Wikidata data structure: the case of an ELTeC edition.

Figure 3 and Figure 4 present examples of ELTeC items in Wikidata. Items include an identifier (framed by red boxes), a list of labels, a description and aliases in different languages (blue boxes), and a list of statements composed of claims (purple), qualifiers (orange) and references (gray). Claims, references and qualifiers are triples, where the predicate (green) is a Wikidata property and the object (yellow) is a value, external URL, date, string or another Wikidata item. Some items have been cropped from the image for clarity.

Every item can be described, as shown in Table 2, by using the Wikidata's unique identifier of a data item QID (e.g. Q107648205, Q559989) as the subject, properties (e.g. P50, P577, P135) and objects, which can be either a literal like "1899" or another item like Q36180, by a series of statements, each providing one fact or information about the item. Table 2 gives several examples of sentences in natural language, their annotation with Wikidata IDs and finally their encoding as Wikidata statements, using reifying for RDF triples (Hernández, Hogan, and Krötzsch 2015) where the same subject is not repeated. Statements with the same subject are separated by a semicolon ";" and the last one is finished by a dot ".".

Ivkova slava is a literary work written by Stevan Sremac.	
Ivkova Slava was published in Belgrade in 1899.	

Ivkova slava (Q107648205) is (P31) a literary work (Q7725634) written by (P50) Stevan Sremac (Q559989).	Q107648205
	P31 Q7725634;
	P50 Q559989;
	P291 Q3711;
	P577 “1899“.

Ivkova Slava (Q107648205) was published in (P291) Belgrade (Q3711) in (P577) “1899“.	
Stevan Sremac was born on 23rd November 1855 in Senta, and he died on 26th august 1906. He was a writer and belonged to realism. He’s VIAF ID is 66526515.	
Stevan Sremac (Q559989) was born on (P569) “23rd November 1855“. in (P19) Senta (Q571136). and he died on (P570) “26th August 1906“. He (Q559989) was (P106) a writer (Q36180) and belonged to (P135) realism (Q667661). He (Q559989) has VIAF ID (P214) “66526515”.	Q559989 P569 “23rd november 1855”; P19 Q571136; P570 “26th august 1906“; P106 Q36180: P135 Q667661; P214 “66526515”.

Table 2. Transforming natural language into Wikidata

4 SrpELTeC Wikidata Entry and Enrichment Automation

Manual population of Wikidata with individual data is often a time-consuming task. As mentioned in Section 1, the initial population of Wikidata with ELTeC editions of novels was done manually, through a user-friendly interface (Figure 3 and Figure 4). In this way 54 novels from the SrpELTeC sub-collection were described in Wikidata by University of Belgrade students through different activities. The control of manual entries revealed that some of the entries were incomplete or contained incorrect information, such as incomplete novel title, an author’s VIAF ID entered as a novel’s VIAF ID , connecting wrong persons as authors (e.g. football player Dušan Đurić (Q116994) instead of the writer with the same name (Q108986248)) or wrong year of the first edition. For these reasons, a systematic validation of manual entries was performed: each dataset retrieved from the headers was compared with the corresponding dataset retrieved by a SPARQL query, in order to identify properties for which statements are missing, yielding results presented in Figure 5. Half of all missing statements were publication places (only 4 items had this information). In some statements the author was missing, which was caused by the fact that these authors were not represented in Wikidata by an QID, so the students used a string label with the author's name instead of the proper author QID. Such problems and shortcomings motivated us to start automating the whole process.

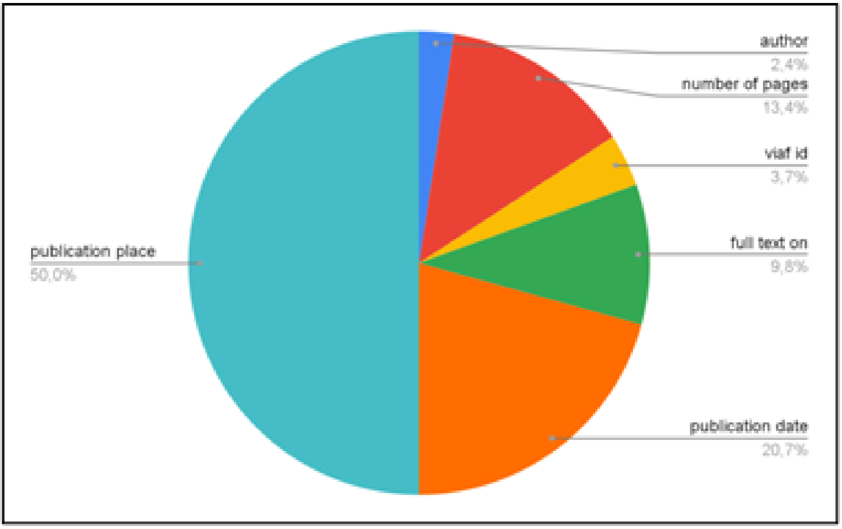


Figure 5. The overview of missing statement per property.

Successful automation of the population of Wikidata for the Infotheca journal (Stanković and Davidović 2021), showed that the population of SrpELTeC Wikidata can also be enhanced by using various procedures and tools, presented in (Turki et al. 2019). The advantage of the ELTeC collection was that the required metadata were available in the header of each novel, as described in Subsection 3.1 The procedure for extraction of all metadata from headers into one *CSV* (comma separated) file in tabular format, appropriate for further transformations, as well as transformation and exploitation of text collections, was integrated in the existing tool for creation, management and exploitation of lexical resources *Leximir* (Ranka et al. 2011).

Before automatically creating items that were missing, it was necessary to fix some problems. According to the WikiProject Books,[21] every book must have a property of either edition (Q3331189) or written work (Q47461344) as (P31). Collaborators working on this project manually added for 48 novels that an instance of (P31) literary work (Q7725634) is also an edition (Q3331189), which was incorrect and we had to remove that from the statement using *QuickStatement* as shown in Table 3 (blue and bold).

Statement	**Remove statement: literary work (Q7725634)**
Glava šećera: ELTeC izdanje (Q106936423) is (P31) literary work (Q7725634)	-Q106936423 P31 **Q7725634**
Bez oca i majke: ELTeC izdanje (Q108838098) is (P31) literary work (Q7725634)	-Q108838098 P31 **Q7725634**

Table 3. Removing a statement from a Wikidata item

Another problem was that students used the property page/s (P304), which represents the location of the claim, for the number of pages, and this had to be replaced by the property *number of pages* (P1104). Also, instead of the property edition (Q3331189) students used the property edition (Q397239), which represents the process of making a version of a work (usually a book). For the number of words a qualifier *determination method* (P459) had to be added in the statement, as illustrated in Table 4 (blue and bold).

Statement	**Add determination method (P459): word count (Q8034324)**
Glava šećera: ELTeC izdanje (Q106936423) number of words (P6570) 11035.	Q106936423 P6570 11035 **P459 Q8034324**
Bez oca i majke: ELTeC izdanje (Q108838098) number of words (P6570) 29321	Q108838098 P6570 29321 **P459 Q8034324**

Table 4. Adding the qualifier in a Wikidata item

[21] Wikidata:WikiProject_Books

For successful entry of novels, the entry of their authors is an indispensable step and a precondition, because the authors have to be separate items in the Wikidata. After checking whether the authors exist in Wikidata by using SPARQL queries, which will be described in Section 5, it was found that 37 out of 68 different authors from the ELTeC collection already existed in Wikidata. For some existing authors some missing properties were detected and added, e.g. P214 (VIAF ID), P21 (gender) and P106 (occupation). For missing authors, items were automatically created, with description and properties, such as gender of the author, date of birth and date of death, which were extracted from the column author. It should be noted that for some authors particular information was missing in metadata (as unknown), for example, there was no VIAF ID, date of birth and/or date of death. For the sake of simplicity, the process of entering authors in Wikidata will not be described, and we will focus on the entry of novels in Wikidata.

For the purpose of our work, it was necessary to add two instances for each novel: the first is a novel as a literary work (Q7725634), which is the subclass of written work (Q47461344), and the second is an edition as instance of (Q3331189). For the edition, two instances are recorded: the first edition (Q10898227) and the electronic, i.e. ELTeC edition (Q59466853). Since we wanted to automate the process of preparing and entering information, we tried different solutions and ended up using two of them, namely, *OpenRefine* and *QuickStatements*. In order to successfully automate the process, several mandatory steps were required. For the first step, the actual preparation of input data, a custom procedure was written that extracts metadata from the TEI file header in tabular form, suitable for further automation, as explained in the Subsection 3.1. Some information for a few novels were missing in TEI headers, because the author was unknown, or the year or the place of the first edition were unknown, while the majority of novels did not have their VIAF ID.

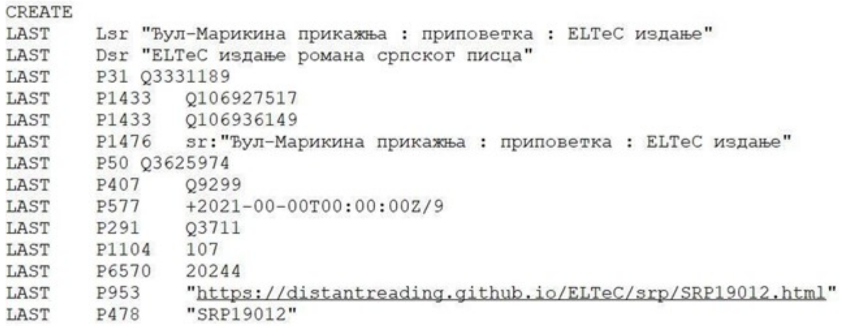
For the novels that were already in the Wikidata and for which some statements were missing, we had to fix all the missing fields. The data were labeled, so that each column represented one statement (predicate) of the novel. It was also necessary to select labels to identify predicates and to create a Wikidata schema. The schema allowed the item to be automatically linked to Wikipedia. Before creating the schema, reconciling each column was necessary. During column reconciliation, a very important process was identification of existing items in Wikidata – a necessary step that enables linking of the file contents to the identifiers (QID) of existing Wikidata items and the creation of new ones for those that do not exist. At this stage, manual verification of data was possible and their correction, if necessary. Each column contained information that was extracted as was presented in Table 1. Some examples of reconciling cells for ELTeC edition of novels are the following:

1. **Title** (P1476) to an entity of type *edition*, *version*, or *translation* (Q3331189)
2. **Author** (P50) to an entity of type *human* (Q5) and then search for match
3. **Language of work or name** (P407) to an entity of type *language* (Q34770)
4. **Number of pages** (P1104) to an entity of type *natural number* (Q21199)
5. **Number of words** (P6570) to an entity of type *natural number* (Q21199)
6. **Published in** (P1433) to an entity of type *text corpus* (Q461183)
7. **VIAF ID** (P214) to an entity of type *VIAF ID* (Q19832964)
8. **Full work available at URL** (P953) to an entity of type *URL* (Q42253)
9. **Publication date** (P577) to an entity of type *calendar year* (Q3186692)
10. **Place of publication** (P291) to an entity of type *city* (Q515)
11. **Volume** (ID of novel) (P478) to an entity of type *volume* (Q1238720)

The next step was editing the Wikidata schema by using OpenRefine. Creating a Wikidata input set schema defines predicates (properties) that will connect subjects and objects in RDF triples. Each statement for a subject has a property and value that can be a Wikidata item, external URL, or literal (string). As presented in Table 1, the property from the first column is related to content (values: items or literals) in the third column. After editing and saving the Wikidata schema it was exported as a *QuickStatements* file. A few lines from this file are given in Figure 6. In the final stage the prepared file was exported in the *QuickStatements* tool and Wikidata items were automatically created.

5 The Overview of SrpELTeC@Wikidata by SPARQL Queries

In this section, we will present a statistical overview of the status of srpELTeC collection in Wikidata, illustrated by characteristic SPARQL queries and their results. We created SPARQL queries for various views, using the



```
CREATE
LAST    Lsr "Ђул-Марикина прикажња : приповетка : ELTeC издање"
LAST    Dsr "ELTeC издање романа српског писца"
LAST    P31 Q3331189
LAST    P1433   Q106927517
LAST    P1433   Q106936149
LAST    P1476   sr:"Ђул-Марикина прикажња : приповетка : ELTeC издање"
LAST    P50 Q3625974
LAST    P407    Q9299
LAST    P577    +2021-00-00T00:00:00Z/9
LAST    P291    Q3711
LAST    P1104   107
LAST    P6570   20244
LAST    P953    "https://distantreading.github.io/ELTeC/srp/SRP19012.html"
LAST    P478    "SRP19012"
```

Figure 6. *QuickStatements* file for creating the statements of a novel.

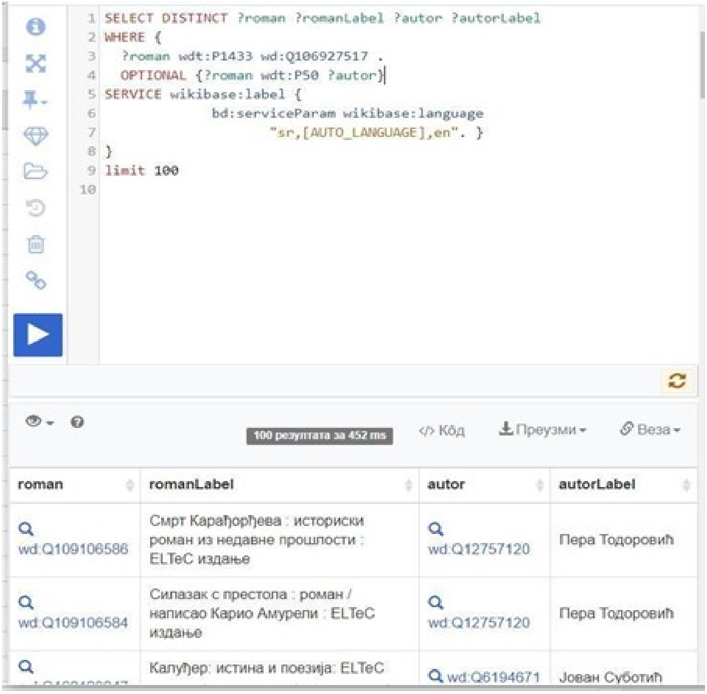
integrated technologies in Wikidata to visualize the results. We wrote queries that retrieved the tables with columns for: the title of the novel, the name of the author, the author's pictures, the year of publication, the authors distribution by gender, etc.

The first validation using SPARQL, retrieved authors that already existed in the Wikidata and it was later used for statistical overview. Before adding Wikidata items, the number of authors in the srpELTeC collection in Wikidata was only 38, while now there are 69 authors, with more than 300 statements as illustrated in Figure 7.

The following query lists authors and novel titles with default view as tree:

```
#defaultView:Tree
SELECT DISTINCT ?author ?authorLabel ?novel ?novelLabel
WHERE {
   # novel published in (P1433) ELteC collection (Q106927517)
   ?novel wdt:P1433 wd:Q106927517;
          # novel instance of (P31) literary work (Q7725634)
          wdt:P31 wd:Q7725634.
   # show the author (P50) of the novel if there is one
   OPTIONAL {?novel wdt:P50 ?author}
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

The statement `?novel P1433 Q106927517` in `WHERE` clause retrieves all novels (?novel) that are published (P1433) in ELTeC: European Literary Text Collection (ELTeC) 1850-1920 (Q106927517). The rows that starts with "#" are comments, introduced to help understand the query. The



```
SELECT DISTINCT ?roman ?romanLabel ?autor ?autorLabel
WHERE {
  ?roman wdt:P1433 wd:Q106927517 .
  OPTIONAL {?roman wdt:P50 ?autor}
SERVICE wikibase:label {
             bd:serviceParam wikibase:language
                   "sr,[AUTO_LANGUAGE],en". }
}
limit 100
```

roman	romanLabel	autor	autorLabel
wd:Q109106586	Смрт Карађорђева : историски роман из недавне прошлости : ELTeC издање	wd:Q12757120	Пера Тодоровић
wd:Q109106584	Силазак с престола : роман / написао Карио Амурели : ELTeC издање	wd:Q12757120	Пера Тодоровић
[illegible]	Калуђер: истина и поезија: ELTeC	wd:Q6194671	Јован Суботић

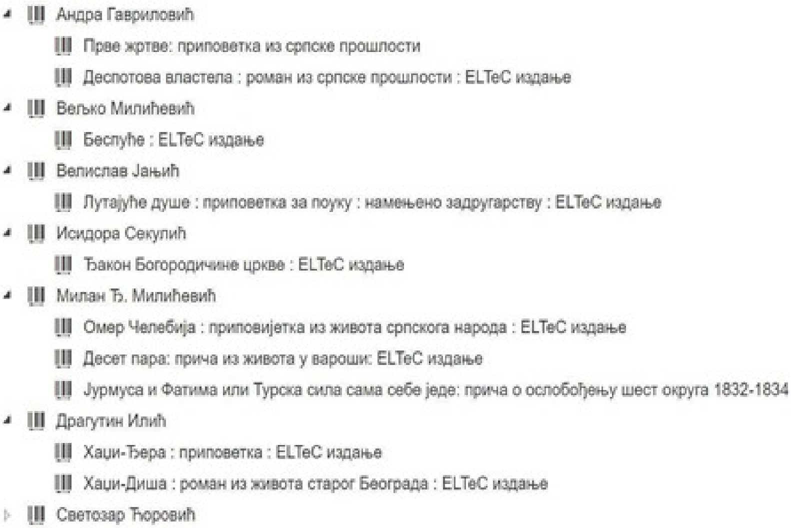
Figure 7. *Wikidata Query Service* with an example.

prefix *wdt:* stands for namespace http://www.wikidata.org/prop/ used for properties and prefix *wd:* is used for objects (QIDs) for namespace https://www.wikidata.org/wiki/. The result of the previous query is given in Figure 8; the whole query and results can be retrieved by Wikidata query service at the link https://w.wiki/4Lja.

The process of entering novels into Wikidata using OpenRefine and QuickStatment was very successful. As a result, there are now 100 novels in Wikidata that are part of the Serbian ELTeC sub-collection and also 10 novels that are in the Serbian extended ELTeC sub-collection, with more than 700 statements.

Using Wikidata Query Service we can display, for example, all novels in the ELTeC collection that have a VIAF ID, number of pages and number of words, with the following query:

```
# defaultView:BubbleChart
SELECT DISTINCT ?novel ?novelLabel ?num_pages ?num_words ?viaf
```



- Андра Гавриловић
  - Прве жртве: приповетка из српске прошлости
  - Деспотова властела : роман из српске прошлости : ELTeC издање
- Вељко Милићевић
  - Беспуће : ELTeC издање
- Велислав Јањић
  - Лутајуће душе : приповетка за поуку : намењено задругарству : ELTeC издање
- Исидора Секулић
  - Ђакон Богородичине цркве : ELTeC издање
- Милан Ђ. Милићевић
  - Омер Челебија : приповијетка из живота српскога народа : ELTeC издање
  - Десет пара: прича из живота у вароши: ELTeC издање
  - Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834
- Драгутин Илић
  - Хаџи-Ђера : приповетка : ELTeC издање
  - Хаџи-Диша : роман из живота старог Београда : ELTeC издање
- Светозар Ћоровић

Figure 8. The graph of authors and their works.

```
WHERE {
  # novel published in (P1433) SrpELTeC coll. (Q106936149)
  ?novel wdt:P1433 wd:Q106936149;
         # number of pages (P1104)
         wdt:P1104 ?num_pages;
         # number of words (P6570)
         wdt:P6570 ?num_words;
         # viaf id (P214)
         wdt:P214 ?viaf.
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

The result of this query is represented in Figure 9, where the size of the circle reflects the number of pages in a novel. Full query results can be retrieved by Wikidata query service on the following link: https://w.wiki/4i9k.

For some novels we imported pictures of cover pages using the Wikimedia commons[22] repository; presently, we are preparing pictures of cover pages for the remaining novels. Figure 10 represents the timeline visualization of

[22] Upload Wizard

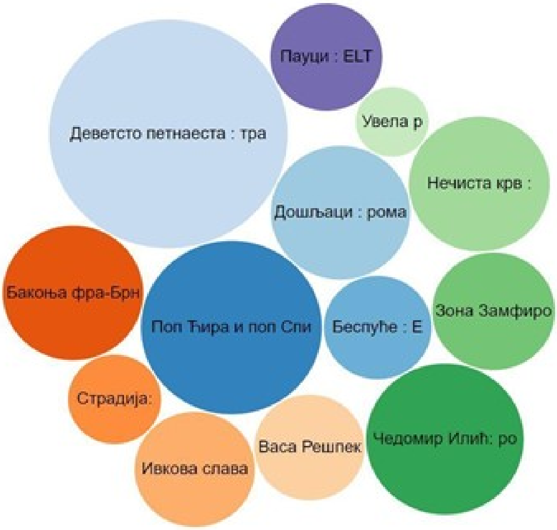


Figure 9. Bubble chart visualization.

novels, sorted by year of their first publication, which was obtained with the following query (https://w.wiki/4LjP):

```
#defaultView:Timeline
SELECT DISTINCT ?novel ?novelLabel ?image ?date ?author
?authorLabel
WHERE {
  # novel published in (P1433) ELTeC collection (Q106927517)
  ?novel wdt:P1433 wd:Q106927517;
         # has edition or translation (P747)
         wdt:P747 ?edition.
         # edition instance of (P31) first edition (Q10898227)
         ?edition wdt:P31 wd:Q10898227;
         # image (P18)
         wdt:P18 ?image.
  # optional date of publication (P577)
  OPTIONAL { ?edition wdt:P577 ?date. }
  # optional author (P50)
  OPTIONAL { ?novel wdt:P50 ?author. }
SERVICE wikibase:label
{bd:serviceParam wikibase:language "sr,[AUTO_LANGUAGE],en".}}
```

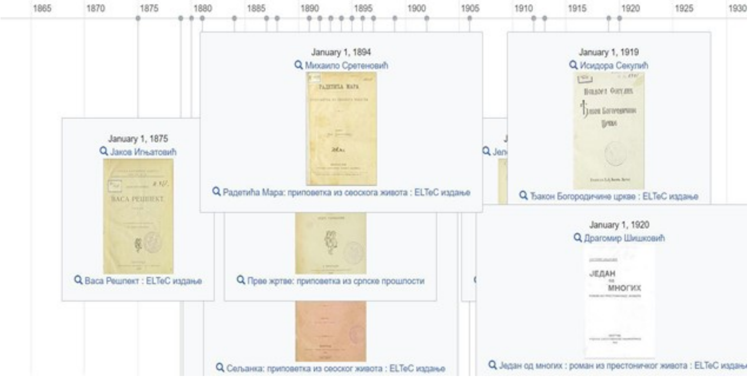

1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
January 1, 1894
Михаило Сретеновић
Радетића Мара: приповетка из сеоскога живота : ELTeC издање
January 1, 1919
Исидора Секулић
Ђакон Богородичине цркве : ELTeC издање
January 1, 1875
Јаков Игњатовић
Васа Решпект : ELTeC издање
Прве жртве: приповетка из српске прошлости
January 1, 1920
Драгомир Шишковић
Један од многих : роман из престоничког живота : ELTeC издањ
Сељанка: приповетка из сеоског живота : ELTeC издање

Figure 10. The timeline visualization.

One of the possible view options is a map preview for queries that have coordinates in the output list. Figure 11 presents a map with places of novel publication for the following query (https://w.wiki/4hSR):

```
# defaultView:Map
# names of the publication places for the first editions
SELECT DISTINCT ?place ?coor
WHERE {
  # edition instance of (P31) first edition (Q10898227)
  ?edition wdt:P31 wd:Q10898227;
         # published in (P143)
         # ELTeC collection (Q106927517)
         wdt:P1433 wd:Q106927517;
         # publication place (P291)
         wdt:P291 ?place.
         # place coordinate location (P625)
         ?place wdt:P625 ?coor.
SERVICE wikibase:label
{bd:serviceParam wikibase:language"sr,[AUTO_LANGUAGE],en".}}
```

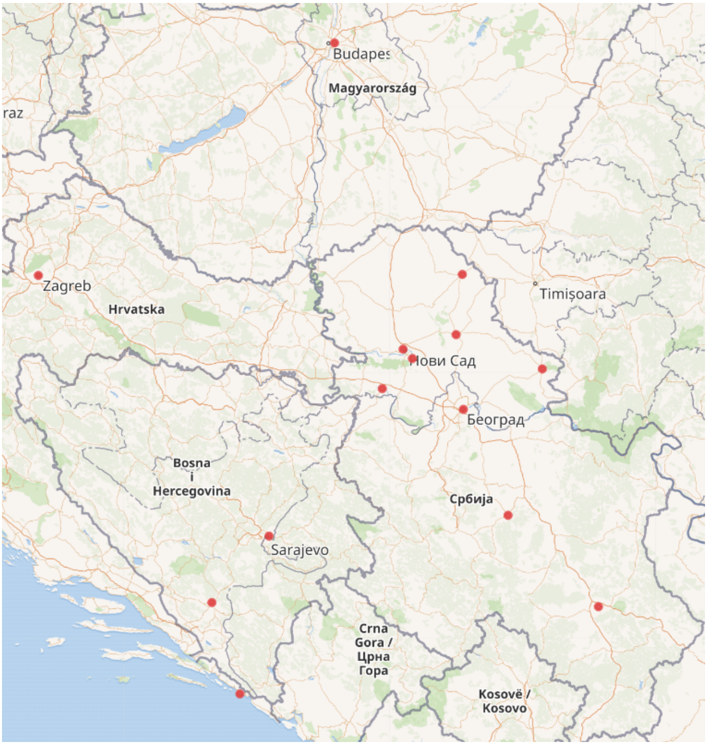


Figure 11. The map with places in which novels were first published.

It is also possible to visualize data as interactive graphs of authors and ELTeC editions (https://w.wiki/4j6D), where the click on an item reveals a set of its properties and related items (Figures 12 and 13).

```
#defaultView:Graph
SELECT DISTINCT ?author ?authorLabel ?edition
?editionLabel
WHERE {
  # published in (P1433) ELTeC collection (Q106927517)
  ?edition wdt:P1433 wd:Q106927517;
         # instance of (P31)
         # version, edition, or translation (Q3331189)
         wdt:P31 wd:Q3331189.
```

```
             # publisher (P123)
             # COST action "Distant Reading for European"
             # Literary History" (CA16204) (Q106926624)
             wdt:P123 wd:Q106926624.
      # optional author (P50)
      OPTIONAL {?edition wdt:P50 ?author}
    SERVICE wikibase:label
    {bd:serviceParam wikibase:language"sr,[AUTO_LANGUAGE],en".}}
```

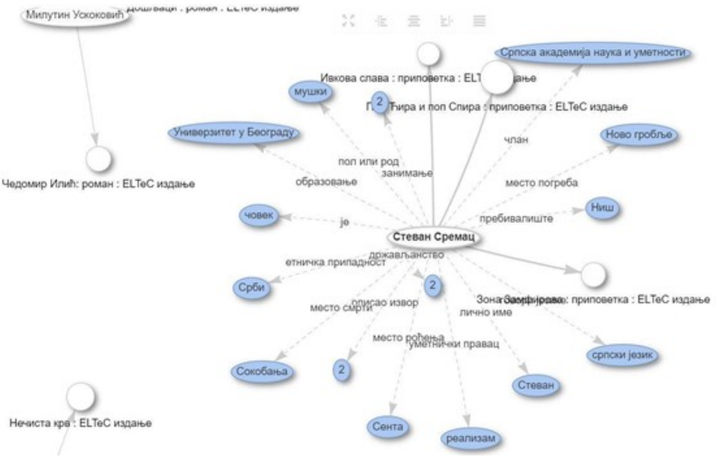


Figure 12. Visualization of an author (Stevan Sremac) and the set of his properties.

As we have already mentioned, the result of this work is that now each novel in Wikidata has items for two editions, the first edition and the electronic (ELTeC edition). Currently, there are 110 ELTeC novels in Wikidata – 100 from SrpELTeC and 10 from SrpELTeC-extended, and since each novel and its associated editions have at least 20 statements, the results is that there are more than 2500 statements for the whole SrpELTeC collection.

6 Conclusions and Future work

This paper presented the automation of the preparation and import of data to Wikidata, illustrated by SrpELTeC, the Serbian sub-collection in the ELTeC multilingual collection (European Literary Text Collection). After the

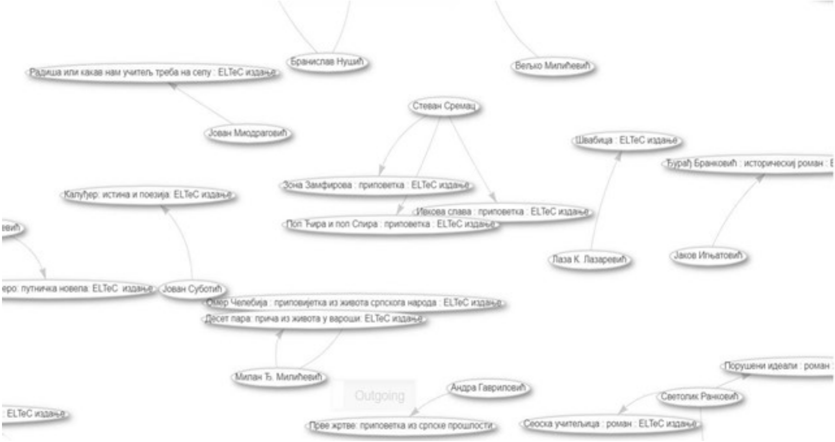

Бранислав Нушић
Вељко Милићевић
Радиша или какав нам учитељ треба на селу : ELTeC издање
Стеван Сремац
Јован Миодраговић
Швабица : ELTeC издање
Ђурађ Бранковић : историческиј роман : E
Зона Замфирова : приповетка : ELTeC издање
Калуђер: истина и поезија: ELTeC издање
Ивкова слава : приповетка : ELTeC издање
Поп Ћира и поп Спира : приповетка : ELTeC издање
евић
Лаза К. Лазаревић
Јаков Игњатовић
еро: путничка новела: ELTeC издање
Јован Суботић
Омер Челебија : приповијетка из живота српскога народа : ELTeC издање
Десет пара: прича из живота у вароши: ELTeC издање
Порушени идеали : роман :
Милан Ђ. Милићевић
Outgoing
Андра Гавриловић
Светолик Ранковић
: ELTeC издање
Прве жртве: приповетка из српске прошлости
Сеоска учитељица : роман : ELTeC издање

Figure 13. The graphs of ELTeC novels and their authors.

extraction of metadata from TEI headers, mapping with Wikidata schema was defined and the synergy of OpenRefine and QuickStatements tools was used for import. As a result of this work, there are now 110 novels from the ELTeC collection in Wikidata, with associated items for the first edition and the electronic ELTeC editions. That means that approximately 2500 statements were automatically added. Future research will use a list of locations, associated with different texts in the corpus, to explore ways to enrich and relate this data to knowledge bases and build a larger context around it. Also, we plan to add data on the main characters in Wikidata, which will include some basic data: gender, profession, whether the character is fictional or not, and if the character is real, a short biography will be entered. With the basic information for each novel (birthplace of author, residence at time of writing, place of publication), one can begin to relate the ELTeC geodata (place of publication and places of narrative) to other time/space coordinates, and consider more detailed mapping visualizations. The analysis of available data about other editions will be explored, as well as other data related to the novel. The research presented is language independent, and the same approach can be used for automation of data import for other ELTeC collections.

Acknowledgment

The text collection preparation is supported by the COST Action 16204 – Distant Reading for European Literary History support. Linked data development is supported by the Wikimedia Serbia "WikiELTeC – Wikidata about old Serbian novels from collection ELTeC". The authors thank the master students of the Department of Library and Information Sciences of the Faculty of Philology, University of Belgrade for their help in filling in the Wikidata with data on the characters from the novels and the places where the action of the novels takes place.

References

Andonovski, Jelena, Branislava Šandrih, and Olivera Kitanović. 2019. "Bilingual lexical extraction based on word alignment for improving corpus search." *The Electronic Library.*

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The semantic web." *Scientific american* 284 (5): 34–43.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020.*

Hernández, Daniel, Aidan Hogan, and Markus Krötzsch. 2015. "Reifying RDF: What works well with wikidata?" *SSWS@ ISWC* 1457:32–47.

Krstev, Cvetana. 2021. "The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata." *Infotheca - Journal for Digital Humanities* 21 (2): 26–42. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.2.

Krstev, Cvetana, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. "Analysis of the first Serbian Literature Corpus of the Late 19th and Early 20th century with the TXM platform." *DH Budapest 2019,* http://elte-dh.hu/wp-content/uploads/%202019/09/DH_BP_2019-Abstract-Booklet.pdf.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. "A system for named entity recognition based on local grammars." *Journal of Logic and Computation* 24 (2): 473–489.

Loesch, Martha Fallahay. 2011. “VIAF (The Virtual International Authority File)–http://viaf. org.” *Technical Services Quarterly* 28 (2): 255–256.

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. “Scholia, scientometrics and Wikidata.” In *European Semantic Web Conference,* 237–259. Springer.

Ranka, Stanković, Obradović Ivan, Krstev Cvetana, and Vitas Duško. 2011. “Production of morphological dictionaries of multi-word units using a multipurpose tool.” In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland,* 77–84.

Shah, Urvi, Tim Finin, Anupam Joshi, R Scott Cost, and James Matfield. 2002. “Information retrieval on the semantic web.” In *Proceedings of the eleventh international conference on Information and knowledge management,* 461–468.

Stanković, Ranka, and Lazar Davidović. 2021. “Infotheca (Q25460443) in Wikidata.” *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. https://doi.org/10.18485/infotheca.2021.21.1.5.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. “Named Entity Recognition for Distant Reading in Several European Literatures.” *DH Budapest 2019.*

TEI Consortium, ed. 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 4.3.0(31-08-2021).* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. “The Serbian Part of the ELTeC Collection – from the Empty List to the 100 Novels Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.1.

Turki, Houcemeddine, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. 2019. “Wikidata: A large-scale collaborative ontological medical database.” *Journal of biomedical informatics* 99:103292.

Vrandečić, Denny, and Markus Krötzsch. 2014. “Wikidata: a free collaborative knowledgebase.” *Communications of the ACM* 57 (10): 78–85.

Андоновски, Јелена. 2019. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки факултет.

From Onions to Champagne – Food and Drink in the SrpELTeC Corpus

UDC 811.163.41’322.2

DOI 10.18485/infotheca.2021.21.2.5

Duško Vitas
vitas@matf.bg.ac.rs
University of Belgrade
Faculty of Mathematics
Belgrade, Serbia

ABSTRACT: The paper presents a set of examples excerpted from the Serbian ELTeC collection, which illustrate eating habits and language about food in the Serbian population from the second half of the 19th and the beginning of the 20th century. The results are confronted with ethnographic sources and description of private life from that time, as well as with the language of food illustrated in few existing cookbooks in Serbian of the time. Unlike these sources, the examples from the corpus show not only what food was used at that time, but also the attitude of the local population towards food, as well as elements of their food taste as part of a collective identity.

KEYWORDS: corpus, language processing, Serbian language, private life, nineteen century, food, ELTeC.

Translated by: Jelena D. Bajić

PAPER SUBMITTED: 2 December 2021
PAPER ACCEPTED: 15 December 2021

1 Introduction

The accounts of private life of Serbs in the past feature, among other things, reports about the eating habits of the population (Фотић 2005). These descriptions, based on the available written resources and archaeology sources, paint a picture of the available and everyday foodstuffs. But the range of food at hand and the dishes that were prepared depended on a wide variety of factors: rural diet differed from its urban counterpart: in the areas dominated by the Turkish influence, the meals were unlike those eaten in the regions where the German influence prevailed; people’s financial situation affected their nutrition, while technological innovations in the food preparation process changed the contents of the menu. A variety of other factors, such as soil composition, local climate or mastery of various agricultural skills had a significant effect on dietary styles. In view of such a complex set of circumstances, the picture about eating habits that can be pieced together from the available historical sources is limited to what was documented, neglecting the intricacies of preparing everyday, ephemeral, quite simple meals of the diverse population about which no traces whatsoever remain.

On the other hand, the testimonies about the taste (or tastes) of the peoples populating the vast area spreading from Thessaloniki to Buda and Szentendre and even farther to Constantinople and Vienna remain unmentioned in the available documents. What is meant here by taste involves not only one's impression about a certain meal, but also an element of the collective identity (Montanari 2011). There is no data as to the kinds of food deemed desirable, or indeed rare, unusual or unacceptable by the members of the different strata of this complex society. One of the examples of the clash between well-established and new tastes in this sense can be found in (SRP18935: *86):[1] Zlata, a harem slave is brought the most expensive Turkish dishes, which she unwillingly accepts, but she frequently thinks about the food common to her village. New food originating from Vienna can also come in conflict with the deep-rooted tastes as in (SRP18941: *85). Moreover, the attitude towards food is not static: it changes continually over time. The meals readily made at one point or those that someone grew accustomed to get eclipsed at another, as a result of a change in the collective taste or dramatic changes in life's circumstances. This phenomenon of the evolution of taste has been spotted as early as in the prefaces to different editions of SRPSKI KUVAR (*Serbian Cookbook*) written by Katarina Popović-Midžina (Поповић-Миџина 1891, 1911), while the act of suppressing the developed taste in the face of cruel refugee hunger is illustrated in the fragment from (SRP19201: *140-1).

The sources lack data on the evolution of the language referring to food and the changes of meaning of certain words. Based on these old written sources, it is easy to get the impression that they describe the same food

[1] The references to certain works forming part of the SrpElTeC corpus follow the (work: number) pattern where work is the code assigned to each novel as listed in the addendum to this issue of *Infotheca*, while number refers to the page number in the novel. A selection of examples illustrating certain items in the text is given in the Appendix at the end of the paper. The examples cited in the text are marked with the symbol * in front of the page number, unless marked otherwise.

and dishes that we are familiar with today. However, both the individual words and the language of food changed as much as the ingredients themselves. The first Serbian cookbooks appearing in the 19th century can be used as sources of examples of the evolution of the language of cooking. A stable nomenclature for certain foodstuffs, ways of preparing food or indeed meals is still non-existent. While our ancestors knew of *tomato* (PARADAJZ) or *frying in breadcrumbs* (POHOVANJE), for example, in the cooking jargon, they are also referred to as *red tomato* (CRVENI PATLIDŽAN) or simply *frying* (PRŽENJE). AJVAR, nowadays a mainly a red pepper dish was termed SERBIAN AJVAR as early as in PATIN KUVAR (*Pata's Cookbook*) (Марковић 1959). It was made of bell peppers, which differed from the then common meaning of the term, since it was considered to be a synonym for *caviar*, etc.

Still, is there a way of capturing at least a fraction of tastes of our ancestors, having in mind the limitations that necessarily make such an attempt but a rough approximation? One of the pathways of inquiry bypassing the historically significant material is made available through the SrpELTeC corpus data. The texts included in this corpus vary as regards the time period in which the story takes place and the type of plot, but in terms of food, they do not go beyond the limits of what readers in the past would have understood and accepted as something edible (SRP18960: *76-7). While the historical novels found in the corpus bring great battles and rebellions from the past back to mind, crime novels construct the plot around crimes, romances describe the emotional life at the time, when it comes to food, the language that describes it was always concrete and precise for the benefit of the then readers. Food parlance of the past can be of interest to the contemporary reader in several ways. On the one hand, it illustrates eating habits of the people living at a certain time, even the social strata about which there are no ethnographic, historical or other records. This is particularly true of the everyday diet of the less well-to-do population in urban areas, about which no reliable data exist. Even in the languages with a long tradition of recording eating habits, the conclusions about common folk food were drawn indirectly, based on what was recorded about alimentation at courts or in monasteries (Flandrin 2002; Montanari 2011). In addition to everyday meals, the corpus contains examples of sophisticated gastronomic knowledge and pleasures and they too failed to come into the focus of attention of ethnographic writings. Nevertheless, there is a risk of all these examples being interpreted through the lens of the contemporary cooking vocabulary, thus creating a false picture about eating habits and making sense of the language of food as it was used at the time.

Cooking handbooks, published in Novi Sad throughout the period are few and far between and they depict primarily European cuisine under strong German influence. The reach of this cuisine extended to the urban areas in what was then known as Serbia too. The handbooks feature instructions for preparing certain meals mentioned in the SrpELTeC corpus. Still, the descriptions found in cookbooks and those focusing on gastronomic experiences are worlds apart. This is convincingly shown in two dictionaries: Ducasse's *Dictionnaire amoureux de la cuisine* (Dikas 2016) and Millau's *Dictionnaire amoureux de la gastronomie* (Mijo 2012). With the former, written by an eminent chef examining the conditions under which to choose the ingredients or compose dishes that would satisfy guests' tastes, but not the personal taste of the chef (or in all probability the guest either), the latter has Millau, a relentless critic of everything that is brought before him in the form of a dish, placing his senses above any culinary authority. Nonetheless, satisfaction during a meal depends especially on one's mood, as shown in the scene from (SRP19100: *203).

The present text is organized in the following way: Section 2 describes the procedure governing the analysis of the SrpELTeC corpus; some of the results obtained by means of this analysis will be presented in Section 3. Special emphasis will be placed on the relation between the examples identified in literary works and the instances from other sources, dating back to the time when corpus material was created. Closing considerations are given in Section 4.

2 Procedure

The analyzed version of the srpEltek corpus consists of 104 novels and longer short stories selected according to the criteria defined by the Distance Reading action team. This version approximately corresponds to the reference corpus, with the exception of certain slight differences stated in the appendix. The corpus also includes the novels left out of the reference version in order to satisfy common criteria for building a corpus in different languages as part of the action (e.g. no more than three novels by the same author).

The analyzed corpus consists of a total of 5,400,000 simple words processed by using Unitex[2] with the help of a system of electronic morphological dictionaries for Serbian.

[2] Multilingual Corpus Processing Suite Unitex/GramLab

The texts forming part of this corpus were analyzed first one by one, separately, and then as a chronologically ordered entity, so as to monitor the distribution of the development of certain linguistic phenomena over time, as well as their collective frequency. Figure 1 provides an example of the results of such a collective analysis showing the distribution of appearance of the different forms of the noun *potato* in the corpus. This figure indicates that the initial 40% of the (chronologically ordered) corpus contains less than 20% of all realizations of this noun and that its frequency increases sharply in the second half of the corpus. Such distribution corresponds, for example, to the observations of foreign travelers about the cultivation and use of potatoes in Serbia in the 19th century(Костић 2019).

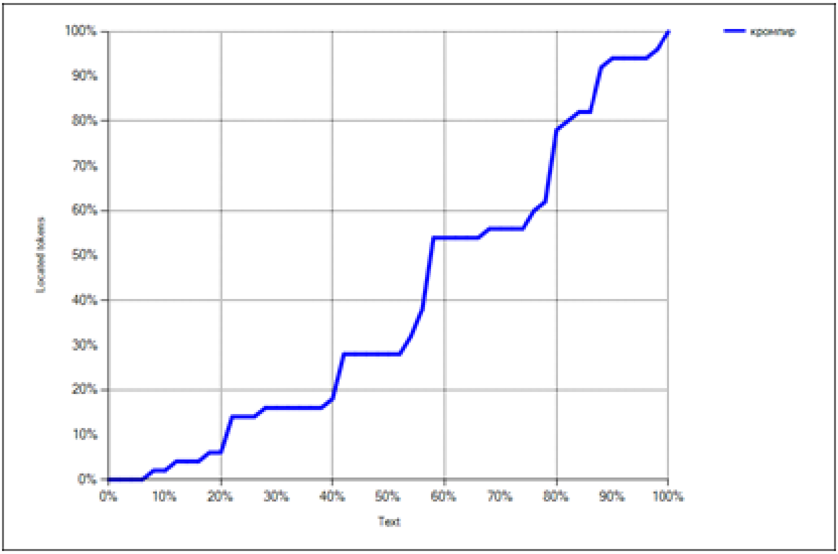


Figure 1. Distribution of the query `<krompir>+<krtola>+<<krump>>`.

Text analysis itself relies on the lexical resources developed for Serbian, especially the semantic markers built into the system of electronic dictionaries described in more detail in (Крстев and Лазић 2015; Krstev et al. 2017; Vitas and Krstev 2012). Besides semantic markers, morphological filters were also used as one of the means of discovering derived words in the records containing the lexemes belonging to the culinary domain. Semantic markers make extracting relevant examples from the corpus possible, but the result must be manually validated due to the presence of homographs. For instance, the string ZOVE may originate from the name of the plant ZOVA *(elder)* or the verb ZVATI (*call*), while the string BLAŽENA may be derived from the adjective BLAŽEN (*blessed*) or be part of the name of the plant BLAŽENI ČKALJ (*Blessed Thistle – Cnicus benedictus*).

The examples extracted via semantic markers show that further development of the dictionary must include the addition of an ontology describing the relations between certain culinary terms. At the existing level of processing, the relation that exists, for instance, between the entries forming the string *wheat – flour – bread* is not explicitly stated. A simulation of this kind of ontology has, in some cases, been done by using local grammars. The formalization of the indicated relation can be achieved, in part, in interaction with the existing lexical resources, but it is an exceptionally complex endeavour, in general.

Variations in the naming of phenomena pose another problem in the process of identifying food parlance in the corpus. Thus, *potato* (KROMPIR) is also KRUMPIR, KRUMPIJER or KRTOLA, *French beans* (BORANIJA) is also BURANIJA and *beans* (PASULJ) can also be referred to as GRA(H), etc.

As the plot in the novels takes place in different cultural settings, the extracted examples should have to be assigned not only temporal markers, but also the markers of dominant cultural influence. However, such markings, although potentially useful are not the primary aim of this analysis. Our goal is to determine what kind of food is mentioned in the corpus and possibly the relation between the attestations identified in this manner and the ethnographic and historical sources referring to the domain of food. In other words, we are, above all, interested in what could be said about food in Serbian at that particular point in time, as well as the nature of the attitude towards food in the past. Brillat-Savarin's statement: "Tell me what you eat and I will tell you what you are" from *Phisiologie du Goût* (1825) has additional significance in this corpus: what is brought to the protagonists of the stories as a meal, defines not just their identity, but also the wider cultural and geographic context of the milieu in which they live.

The search for culinary terms in the corpus was performed across sections consisting of entries sharing thematic similarities, so that the results obtained from the corpus could be compared to other sources. The comparison was made relative to the relevant historical and ethnographic sources. We focused our attention on the kinds of food or dishes whose past meaning or use differ from those opted for today. Because of the space constraints imposed by this article, the present review could not be comprehensive.

3 Results

The language of food is featured throughout SrpELTeC in a variety of ways, sometimes only in passing and sometimes the entire sumptuous feasts with the accompanying rituals were described, as in (SRP18950). Such examples that could be recognized by using the available tools were singled out and classified. The classification made it possible to organize these examples into sections featuring entries that share thematic similarities, thus giving an insight into certain aspects of the process of preparing and serving, as well as enjoying food at the time. We will first examine the range of foodstuffs that can be used to prepare a meal, subsequently focusing on some dishes and beverages that found their way to tables of the protagonists of the literary works forming the corpus.

3.1 Food of Plant Origin

Food originating from plants can be classified into several groups. The first group is constituted by cereals. The word *cereal* (ŽITARICA)[3] is not present in the corpus. Instead, it is substituted by *grain* (ŽITO) and its many meanings. Grain can refer not only to wheat and corn, but also other cereals, as in (SRP19024: 147), for example, or maybe in (SRP19102: 310) and (SRP19132: 216). Individual cereals are also mentioned, especially *corn* (KUKURUZ) [236][4] (in (SRP18992:45) – KURUZA or in (SRP18970: 33) - *corn fields* (MURUZNE NJIVE), *wheat* (PŠENICA) [52] (or ŠENICA [15]), *oats* (OVAS or ZOB) [50], *barley* (JEČAM) [28], *rice* (PIRINAČ (PIRINADŽ)) [19], *broomcorn* (SIRAK) [2], *rye* (RAŽ) [1], *buckwheat* (HELJDA) [1] <36>.[5] *Millet* (PROSO) appears only once, indirectly, as *millet bread* – PROSENICA <37>.

Flour [108] and subsequently baked goods are made of cereals: *bread* (HLEB (HLEBAC, LEB, LJEB, LEBAC)), *round unleavened bread* (POGAČA), *flat*

[3] In Vuk Stefanović Karadžić's 1852 *Serbian dictionary* (SRPSKI RJEČNIK), the only meaning given for the word ŽITARICA is "grain transporting ship", while the corpus features the word BOLOZANKA in (SRP18620:7) with the same meaning.

[4] The number shown in square brackets [...] refers to either the frequency of a lemma or the one obtained from a lemma (that includes the derived forms) by applying a morphological filter. For example, the filter <<^pšenič>> extracts the examples for the noun PŠENICA and the adjective PŠENIČNI.

[5] Ordinal numbers referring to the selected examples of use of the terms denoting cereals and vegetables given in the Appendix as subsection *Examples of Food of Plant Origin* appear in the text in brackets < i >.

unleavened bread (LEPINJA), *round loaf of wheat bread* (SOMUN), *round bread* (SIMIT), *rusk* (PEKSIMIT), *commercially manufactured biscuits* (BEŠKOT) (a total of over 800 mentions), as well as *corn bread* (PROJA (and PROHA)) [62], along with *rolls* (KIFLE), *buns* (ZEMIČKE), *bagels* (ĐEVRECI), *pretzels* (PERECE)...

How does this data measure up to the historical and ethnographic sources? Карић (1887, 109) states that "corn bread and leavened and unleavened bread made of wheat, barley, oat and buckwheat flour is the main food of our peasants in the lent season" which corresponds to the data obtained from the corpus. The following quote from (Мијатовић 1908) reads that in the Levač region "rye (...), oats, buckwheat and millet (fine grained corn) are rarely sown". Besides, *spelt* (KRUPNIK) and *meslin* (SURAŽICA) that are absent from the corpus are mentioned in (Зиројевић 2005). These sources fail to make reference to a wide variety of baked goods.

At first sight, the selection of vegetables does not differ much from what is available in the open-air markets today. The ethnographic sources describing eating habits of the time reference varying selection of vegetables depending on author. Thus, Карић (1887, 110) mentions *beans* (PASULJ), (extremely hot) *peppers* (PAPRIKA), *cabbage* (KUPUS), *broad beans* (BOB), *French beans* (BORANIJA), *lentils* (SOČIVO), *potatoes* (KROMPIR), *radish* (ROTKVA), *rice* (PIRINAČ), *dock* (ZELJE), *cucumbers* (KRASTAVAC), different kinds of bulbous vegetables (*garlic* (BELI LUK), *onion* (CRNI LUK), *leek* (PRAZILUK), *chives* (VLAŠAC), *bunching onion* (ALJMA)). This list is extended in (Мијатовић 1908) to include *squash* (TIKVA), *mushrooms* (PEČURKE), *bell peepers* (BABURE), *pickling peppers* (TURŠIJARE), RED TOMATOES (CRVENI PATLIDŽAN), whereas *nettle* (KOPRIVA), *green amaranth* (ŠTIRA) and *white horehound* (PEPELJUGA) are used in the same way as *dock*. The author also lists little known vegetable crops that are not cultivated; these are: *peas* (GRAŠAK), *eggplants* (PLAVI PATLIDŽAN, MODRI PATLIDŽAN), *lentils* (SOČIVO), *parsley* (PERŠUN), and *dill* (MIROĐIJA). He also mentions *beet* (CVEKLA), describing it as "the root of *patience dock* (PITOMO ZELJE)" – thus, could be no different from *chard* (BLITVA).

The most comprehensive catalogue of edible plants is found in (Зиројевић 2005) where the above-mentioned selection is expanded by adding *carrots* (REPA (ŽUTA REPA, ŠARGAREPA and MRKVA)), *beans* (GRAH), *spinach* (SPANAĆ), *sorrel* (KISELJAK), *eggplant* (PLAVI PATLIDŽAN), *kohlrabi* (KELERABA), *cauliflower* (KARFIOL), *artichoke* (ARTIČOKA), *asparagus* (ŠPARGLA), *pumpkin* (BUNDEVA), *celery* (CELER), *kale* (KELJ), *horseradish* (REN) and *zucchini* (TIKVICA). The differences between various cultivated vegetable crops encountered in diverse sources stem from the fact that older written works provide descriptions related to Serbian 19th century countryside, disregarding the range of vegetables available in towns, especially in Austria-Hungary, where the selection of vegetables included in the diet was considerably richer. Hence, *onions*, *garlic*, *cabbage*, *potatoes*, *kohlrabi*, *rice*, *beans*, *string bean* (MAHUNE), *green beans* (ZELENI PASULJ) and *French beans*, *peppers* and *bell peppers*, *eggplants*, *tomatoes*, *cucumbers*, *kale*, *pumpkins*, *squash* and *gourd* (JURGET), *peas*, *spinach*, *asparagus*, *dock*, *chives*, *celery*, *carrots*, *truffles* (GOMOLJIKA), *mushrooms*, *button mushrooms* (ŠAMPINJONI), *parsley*, *dill*, *parsnip* (PAŠTRNAK)… are named in (Поповић-Мицина 1891) where the following remark is given about bell peppers: "Take bell peppers, Bulgarian ones, nice ones…"

The corpus attests to the presence of almost all these vegetable crops. First and foremost, there are general terms, such as *vegetables* (POVRĆE) [28], *greens* (ZELEN) [7] and *greenstuff* (ZELJE) [10] but especially – *stewed vegetables* (VARIVO) [5] mostly referring to *pulses* (MAHUNARKE). An example of their occurrence can be found in Appendix 2 under items <1, 2, 18>. The examples of cultivation of individual kinds of vegetables have also been confirmed, primarily bulbous vegetables, above all *garlic* and *onions* [156]. Other vegetables of the bulbous variety, namely, *leek* [4], *pichling onion* (ARPADŽIK) [1] are rarely found, while *chives* and *bunching onions* are absent <3, 4, 15>. These are followed by *beans* (GRAH, GRA') [97] <5, 15>, *French beans* (BURANIJA) [5] <6> and *lentils* (SOČIVO) [10] <7>, with the form LEĆA missing from the corpus. Other kinds of pulses on record include *peas* [3], *broad beans* [8] and *chickpeas* (LEBLEBIJE or NAUT) [13] <8, 9, 10>. Furthermore, both *cabbage* and *sauerkraut* (KISELI KUPUS) [72] <11> were confirmed as listed in the corpus, together with *peppers* [56] <12, 13[6]>, *potatoes* (KRTOLA, KRUMPIR, KRUMPIJER) [48] <14, 15>, *cucumbers* (KRASTAVAC, KRASTAVICA) [31] <16, 17>, *radishes* [10] and *red radishes* (ROTKVICE) [2] <19>, *asparagus* [5], *cauliflower* [4] and *kale* [2]. *Pumpkin* [17], *squash* [44] and their variant names: DULEK [7] and ŠUĆURKA [1], JURGET [1] and LUDAJA [2] <23, 24, 25> are also present. Cultivated plants include *horseradish* [10] <26>, *spinach* [5] <27>, *lettuce* (salata) [5] <28>, *parsley* (MAGDANOS) [3] <4, 5>, *celery* (ĆERVIZ) [3] <4, 29>.

[6] Example <13> reveals why *bell peppers* in (Поповић-Мицина 1891) are referred to as Bulgarian.

There are red tomatoes and there are eggplants[7] [11] <13, 14> but the examples <30> show that both vegetable crops were regarded as foreign elements of the cuisine by the residents of the Belgrade Dorćol quarter in the late 19th century. The name *tomato* [4] (CRVENI PATLIDŽAN) appears in the corpus at that time among the Serbs in Vojvodina <31> and somewhat later in the town of Niš as CRVENI FRENC. [32]. Such late appearance of tomatoes corresponds to the data provided by (Zirojević 2019).

The examples referring to other vegetables cited in (Поповић-Мицина 1891) (Popović, 1891) indicate that they are not widely present in the diet of the time. *Beet* (CVEKLA (PANDŽAR)) [6] is mentioned because of its colour or as cattle feed. *Carrots* are known as *yellow radish* (ŽUTA REPA) [1] <33> and just like beet, they get mentioned because of their color, rather than as an element of the diet.[8] The fact that neither beets nor carrots were common is made evident by their very names, taken from Hungarian, while PANDŽAR is of Turkish origin.

Mushrooms [12] are used only metaphorically, as in the expression mushrooms after rain, while *okra* (BAMIJA),[9] *bell peppers*, and *pickling peppers* are not mentioned at all. *Sunflower* [21] <34, 35> is known as a decorative plant only.

Other food of plant origin, namely, fruits, spices and medicinal herbs will not be mentioned here.

The frequency of the above-mentioned plants confirms their presence in the diet, providing hints about the nature of the common meal, as well as about forbidden or undesirable food, at the same time.

Onions, whose frequency is high are often the only ingredient of meagre meals, in addition to bread (see the examples in the Appendix, Subsection *Examples of the Simplest Meals - Bread and Onions*). This meal structure corresponds to the description given in (Карић 1887, 109) and (Мијатовић 1908, 12). A variety of vegetables makes a more varied and richer diet possible, illustrated, for example, by Gican's feasts in (SRP18741:117-21) or New Year's Eve supper in (SRP18911: 65) (see Subsection *A Description of a Feast* in the Appendix). Both descriptions show

[7] In Serbian, the same noun (PATLIDŽAN) is used for both *tomato* and *eggplant* that are distinguished by their respective adjectival modifiers, *crveni* (RED) and *blue*, *dark* (PLAVI, MODRI).

[8] Поповић-Мицина (1891) (Popović, 1891) cites recipes for preparing *carrots* and *beet*.

[9] BAMIJA might be synonymous with the term BABNJA in (SRP19012: 27).

that gourmet cuisine of townspeople, where onions were no longer key, coexisted with the modest menu of peasants and workers.

In the case of some of the terms, referring to certain plants, it is unclear what exactly is meant by them. For instance, ZELJE, according to (Белић 1959–2021) (RSANU) can be understood both as a general term for plants and taken to mean "different herbaceous plants (...) whose leaves are often eaten". In the example (SRP18922: 118) "GRGO IZNESE ZELJA I SUHE RIBE" ("*Grgo brought (herbaceous) plants and dried fish*"), ZELJE is as undefined as the kind of fish brought to the table. Nevertheless, even as regards the plants whose names are unambiguous, we cannot be sure whether they are the same as their counterparts today. We can find out something about their appearance from the metaphors related to their characteristics: *oranges like golden apples* (SRP18910:75), *small like a cucumber* (SRP19203:110), *pale like a lemon* (SRP18760: 351), *yellow like a lemon* (SRP19070: 92), *pale and blue like an eggplant* (SRP19021:67), *all red like a pepper* (SRP18892:240), *as hot as a pepper* (SRP19001: 25) (see also (Krstev 2021) in the same issue).

Some plants used as food are not mentioned. For instance, *okra* (BAMIJA), referred to by Vuk Stefanović Karadžić in his *Dictionary* is not mentioned even once in (Грђић-Бјелокосић 1908), while (Марковић 1959) says the following, specifying that *okra* is an indispensable ingredient of stews: "In the past you could see heaps of this vegetable in Belgrade's open-air market, but it is rare today. Okra is a very pleasant herbaceous plant in the summer and many people like it better than any other". Did okra disappear from stews because of its origin that was considered to be too oriental? *Bulgur*, which according to (Трojановић 1983, 319) is used and very valued by peasants, who use it instead of rice, suffered a similar fate. It appears in the corpus only once in the form of BUNGUR (SRP19030: 122). *Rice*, whose use is mainly associated with Turkish cuisine owes the majority of its appearances in the corpus to the works published after the year 1900 and mainly in relation to oriental dishes. This shows that taste, as part of the collective identity, as far as the choice of foodstuffs is concerned was determined by the reasons that could be different from the objective nutritional value of food.

3.2 Food Staples or Groceries

What information does the corpus provide regarding food staples that cannot be produced in the countryside, such as oil, sugar or coffee?

Salt was recorded 107 times, 40 times of which in the expression HLEB I SO (*bread* and *salt*) (including the variations of the lexeme HLEB) as a symbol of hospitality. Twenty instances are related to the purchase of salt in a store or its delivery from Vidin, for example (SRP18790:40), or to Zlatibor from Belgrade (SRP18880: 55). In all probability, the purchase of salt was an important expense, since it is linked to the buying of *peasant shoes* (opanci) no less than four times (SRP18751: ?), (SRP18992: 45; 70), (SRP18993: 9). It is noted in (Карић 1887, 114) that too much salt is added to food, while the remark that such eating practices should be changed is found in (SRP19102: *382).

Zejtin [53] and ulje (*oil*)[10] [32] appear side by side, sometimes even in the same sentence (SRP18880: 17). In the vast majority of examples, the appearance of oil is related to icon lamps, but it is also used as an element of nutrition, as in (SRP19000: 83) or (SRP19201: 19). It is also necessary when seasoning a salad, as in (SRP18730: 126): „evo večere, dajte samo octa i ulja“[11] (“*here's supper, just get some vinegar and oil*”). According to (Поповић-Мицина 1891), oil is used primarily for seasoning and rarely for frying, in which case, mostly for frying fish.

Oil is among the rare foodstuffs that have to be purchased (SRP19102: 97). A description of a grocery store where oil is sold together with salt and other household items is given in (SRP19140: 178). In villages, it was sometimes distributed by hired coachmen, as well (SRP19101: 137).

The origin of oil usually is not stated. However, based on some examples, it can be concluded that what is referred to is first and foremost olive oil. Thus, *fine oil from Ulcinj* is mentioned in (SRP18590: 541), while olives are linked to oil in (SRP18892: 332) or (SRP19061: 189). Still, other examples indicate that zejtin and in fewer cases ulje is not exclusively olive oil. In (SRP19080: 44) there is the following comparison: poznata žilavka žute boje, jasna i čista kao zejtin (*the famous yellow coloured (wine) Žilavka, clear and pure like oil*). Because of the colour, this oil may not have been of the olive variety. In (SRP19091: 75) it is stated that Anđa noću razgreva zejtin (*Andja heats up oil at night*) and that is why it might have been a different kind of fat. In (SRP18880: 36) med je lagan i tečan kao etirno ulje (*honey is light and liquid like ether oil*). The quality of oil (zejtin) can differ, depending on the degree of refinement: from the example in (SRP19060: 434) šalje u crkvu najprostiji zejtin

[10] Zejtin [53] and ulje [32] are synonyms, both meaning *oil*, just like sirće and ocat below referring to *vinegar*.

[11] *Vinegar* (sirće) [30] or somewhat rarely, ocat [9] is used along with oil. Vinegar can be made from plums, wine, apples or even roses as in (SRP18920: 213).

(*sends oil of the worst quality to church*), to the one in (SRP18941: 64) where NAJČISTIJI BARABANC-ZEJTIN (*the purest Bărăbanț oil*) is used. According to (Зиројевић 2005), oil could also be made from sesame or poppy seeds, "šarlagan", but these plants are absent from the corpus. Linseed might also have been used, but it is mentioned only alongside hemp.

The first appearance of sugar [145] is recorded in (SRP18631: 74) and subsequently it spread evenly throughout the remaining part of the corpus. In (SRP18751), there are 32 instances in the form of loaf of sugar and this particular combination appears only once more in (SRP19061: 10), while in (Тројановић 1983, 289) we find that "SE PAKUJE U HARTIJU [...] ONU PLAVU SA GLAVE ŠEĆERA" ("*it is wrapped in paper [...] the blue one removed from a loaf of sugar*"). Sugar was either imported or produced locally from sugar beet in the late 19th century (SRP19100: 202). It was sold in shops (SRP18750: 40), (SRP18760: 320) on both banks of the Sava and the Danube where it was stored in barrels or sacks. It was served with brandy (SRP18960: 108), coffee or water, usually as a *piece of sugar* (PARČE ŠEĆERA). It was used for making sweets, both oriental ones and those originating from Vienna (SRP19030: 118).

Finally, *coffee* [798], which gets mentioned more times than any other kind of food is indispensable. It is attested throughout the corpus in different forms (KAFA [394], KAVA [349], KAHVA [55]) also KAJMAKLIJA [6] and even [2] *black broth* (crna čorba) e.g. in (SRP18940: 151). In the early ethnographic writings, no reference is made to coffee.[12] Coffee is essential, but it is viewed as a source of evil as early as in (SRP18630: 215) if it is drunk by women. Nevertheless, despite such remarks, the frequency of its occurrence shows that coffee was already an obligatory everyday beverage at the time. Moreover, there were people addicted to good coffee (SRP19193: *20). Precise instructions for making coffee are given in (Драгановић 1855), while (Поповић-Мицина 1891) provides a detailed description of coffee varieties classified by origin and method of preparation, including even a description of a percolator. Coffee arrived from Brazil (SRP19140:168). Coffee beans were unroasted, so they had to be roasted and ground after purchase. Black coffee or cafe au lait was drunk at home, in cafes and also at work where it was made by attendants.

The making of coffee, that is GORKA KAVA, NAJMILIJE PIĆE TURSKO (*black coffee, a favourite Turkish drink*) was described as early as in (SRP18631: 70). At the same time, in (SRP18691: 100), coffee is the morning

[12] ALOVINA, a beverage made of oats or barley, drunk instead of coffee is mentioned in (Тројановић 1983) but this word is not featured in the corpus.

drink of bishops. It was drunk in towns and in villages too (SRP18891: 116). Coffee was made the Turkish way, in a Turkish coffeepot (DŽEZVA) placed on a brazier (MANGAL) (SRP19070: 31) and in a coffee roasting tile stove (KAVE-ODŽAK) (SRP18790: 49). It was served not only in *narrow-necked copper vessels, with a cover used as coffeepots* (IBRIK) and Turkish coffee cups (FILDŽAN), but also in the European manner in a porcelain coffeepot and cups (SRP18892: 294). Coffee was sipped very hot [15] and often served with brandy, wine or *fruit preserves* (SLATKO) [46].

Examples for this section are given in the Appendix, subsection *Examples of Food Staples or Groceries*.

3.3 Dishes

Which dishes were prepared in the past and what were our ancestors able to cook using the above-mentioned foodstuffs? The answers to these questions depend not only on the already well-established tastes, but certainly on cooking skills and the available kitchen tools, as well. Some dishes can be cooked in a copper cauldron hanging over the hearth, others in the oven, in a pot or indeed using a wide array of kitchen tools. Culinary skills are derived either from inherited food preparation practices or the innovations resulting from new knowledge adopted from other cultures. The first cookbook in Serbian (Драгановић 1855) states in its very title that it features recipes collected from German books on cooking. However, it became the basis of local feasts as early as in (SRP18941:385) where Jerotej's theory got transformed into madam Sida's cooking practice.

Meals ranged from the primitive eating using a shared spoon or fingers to pick food from the dining table (SRP18740: *133) to food excesses typical of feasts (SRP18967: *?). While in the former case food was brought or poured on the table all at the same time, in the latter, dishes followed one after the other in a predefined succession throughout the meal. The order of dishes served as part of a meal can be fixed (SRP18741: 121), (SRP18960: 116), (SRP18961: 26),[13] in some settings, even the menu can be permanent and organized by the days of the week (SRP18880: 61).

Were the meals brought to the table in the past, such as soups and broths, beans, moussaka, stuffed peppers, stuffed sauerkraut leaves, stewed sauerkraut, goulash, steaks prepared in different ways and certainly barbecue, ajvar, and other salads the same ones that constitute "local cuisine"

[13] The order of dishes during feasts held by Serbs, Russians, the English, the French and Swedes was described as early as in (Поповић-Мицина 1891)

today (Витас 2018)?[14] Can the examples featuring these dishes, which are the basis of the usual and even national cuisine nowadays be found in the corpus?

However, the names of dishes alone are usually not enough – it is necessary to cross-reference their names with the recipes, if any, dating back to the period in question. In addition to Jerotej's (Драгановић 1855) cookbook, there was also the one by Поповић-Миџина (1891) that had four editions by the year 1920, as well as the 1922 *Cookbook* by Мирковић (1922). All these books on cooking have a shared denominator "srpski kuvar" (Serbian Cookbook), found in their respective titles, despite having been published in Novi Sad. Consequently, they feature approximations of the dishes prepared under strong German and Hungarian influence at the time, while making a modest contribution to the description of the dishes existing on the opposite riverbanks of the Sava and the Danube.

We will look into some of the above-mentioned dishes comparing them to the data attested in the corpus. The aim is to examine whether the name is all that has remained to this day or the dish itself has been preserved in the form cited in the cookbooks of the time. Some dishes have obviously survived to this day, including beans without any animal fat added, beans with bacon or smoked meat, or pap, stuffed sauerkraut leaves or wine leaves, sauerkraut stew with turkey. *Goulash* (GULAŠ, GULJAŠ), however is mentioned only twice in (SRP18630). On the other hand, the somewhat forgotten chicken or lamb stew is mentioned as many as 19 times. Stuffed peppers are absent and moussaka appears only once. Just like other stew-like southern dishes (ĐUVEČE, JANIJA and PAPAZJANIJA), they are rejected in certain regions as completely unacceptable foreign dishes (SRP18880: 72).

Soup and Broth

Nowadays, the difference between a *soup* (SUPA) and a *broth* (ČORBA) is clear: a broth must be garnished with browned flour. Soup is non-existent in the books on cooking until 1920. The only thing discussed are broths. Jerotej provides recipes for around twenty broths, both those with meat and/or fat and those without, but the term *soup* is not present. In Поповић-Миџина (1891), a single instance of the word *soup* (SUPA) is found. Namely, it forms part of the transcription of the German term BRAUNE SUPPE: BRAUNE SUPE. The local equivalent for this in Поповић-Миџина (1891) Midžina is

[14] A similar selection of local meals can be found at Wikipedia article on Serbian cuisine.

dark broth (MRKA ČORBA), the obligatory ingredients of which are hollow bone and beef liver. This dish is referred to as *beef broth* (*soup*) in a recipe appearing as late as in Sofija Mirković's cookbook.

The word *broth* appears in the corpus 75 times, for the first time in (SRP18740: 137) and (SRP18741: 95) where *black broth* (CRNA ČORBA) with grated Parmesan cheese is also found (SRP18741: *121), which is probably the same as Midžina's *Braune Suppe*. A broth can be meat/fat free (SRP19012: 3), or otherwise contain (fatty) beef (SRP18760: *377), lamb (SRP18871: *41), pork (SRP18964: *50), chicken (SRP19102: *65), or other poultry. There is also *fish* (RIBLJA) or *fisherman's* (ALASKA) *broth* (SRP18950: 54), flavoured with kaymak and eggs SRP19102: *65), eggs being an important ingredient (SRP18940: *54). *Sour broth* (KISELA ČORBA) [11] is a true favourite. It is described as containing chicken in (Драгановић 1855), while according to (Поповић-Мицина 1891) any kind of meat can be added to it. The sour quality is obtained by adding vinegar or lemon.

Soup (SUPA) [30] appears for the first time in (SRP18630: 148), therefore, before the first appearance of *broth* (ČORBA), just after the publication of the first edition of Jerotej's cookbook. This can be explained by Jerotej's (Драгановић 1855) and later Midžina's (Поповић-Мицина 1891) insistence on the use of Serbian names of dishes: they rejected the term SUPA on account of its foreign origin and replaced it by the Serbian counterpart – ČORBA. What must be kept in mind here is the fact that fresh beef is used when cooking beef soup; thus, there is the requirement of buying the ingredient at a butcher's shop, which was feasible only in urban areas at the time.

Unlike broths, the composition of soups is not mentioned except in (SRP19001: *27). Still, some people regarded soup as a new dish, since it is stated in (SRP18961: *26) that broth, referred to as soup, is brought to the table, having completely replaced the traditional sour broth. An indication of making a clear distinction between a broth and a soup is found as late as in (SRP19140: *269).

Examples are given in Section *Soup and Broth Examples* in the Appendix.

Kebab (ĆEVAP) and *Grilled Minced-Meat Finger* (ĆEVAPČIĆ), Roasting and Frying

The lexeme ĆEVAPČIĆ (*grilled minced-meat finger*) has been widely replaced by the lexeme ĆEVAP (*kebab*) today (Витас 2018). Before this latest change of meaning took place, ĆEVAP used to refer to pieces of meat prepared or cooked most often on a grill, while ĆEVAPČIĆ meant “finely chopped [...] meat prepared in the shape of small, short sausages and cooked on a grill” (Петровић 1937). In (Поповић-Мицина 1891) only ĆEVAP is found; it is prepared using big chunks of meat or fish that are roasted on a skewer, in the oven or steamed (in water or beef broth). Ćevapčić, however, is absent.

Both lexemes: ćevap [12] and ćevapčić [10] are featured in the corpus. Ćevap, appears for the first time in (SRP18882: *37), and with the exception of an example in (SRP19140: *269) where *kebab* (ćevap) is steamed (in water or beef broth), all other examples feature kebab roasted on a skewer. As early as in (Марковић 1959) (Marković, 1959) it is said that skewered kebab is “a classic dish served for lunch in the field” but that it must not be “three paces long”, that is not exceed half a metre. In (SRP18934: *88) there are words of praise for an arşın (approximately 3/4 metre) long kebab.

Ćevapčić appeared for the first time in (SRP18690: *5). In (SRP19100: 67), their number is reduced to today’s restaurant potion of ten minced-meat fingers.

In the fourth edition of (Поповић-Мицина 1911), grilled minced-meat fingers and *skewers* (ražnjići) are among the dishes that are mentioned for the first time as “Serbian food” in addition to kebab, whose meaning is as described above. A skewer is described there as a small kebab made of walnut-sized pieces of meat, its dimensions being similar to those of a contemporary skewer. Grilled minced-meat fingers were made of meat cut into small pieces (meat grinders would appear later). They were about a finger long. When grilled, both minced-meat fingers and skewers got sprinkled with finely chopped onions. The then version of kebab, therefore is more akin to what is sold today as Greek gyros, but positioned horizontally when roasted, or resembling a giant skewer. *Grilled minced meat patties* (pljeskavice) are nowhere to be found and a hint of what is known today as grilled minced meat patties topped with melted kaymak can be found in (Мијатовић 1908). In other words, the pride and joy of the local cuisine and an indispensable feature of the contemporary restaurant and fast food joint menus experienced considerable changes over the last hundred years, both in terms of its lexis and probably taste, as well.

In addition to barbecue, meat [128] roasted on a spit (SRP19000: *172) is an indispensable element of feasts and other meals, particularly restaurant ones. Like broths, *roasts* (pečenje) vary from roasted pork, lamb, poultry, veal, ox, rabbit, pheasant and even badger. Roasts are eaten with fingers, according to the Turkish custom (SRP18950: 125), (SRP18760: 338) or get carved up (SRP18620: 37) and put in bowls or plates to be carried to the table or stored in baskets prepared for picnics. Another belief regarding traditional dishes is brought into question here: pork roast is as common as lamb roast. It is served warm or cold and it can be bought in restaurants. An interesting remark about the quality of roasted meat in roadside restaurants can be found in (SRP19001: 237).

Besides meat roasted on a spit, exceptionally rare occurrences of individual pieces of meat fried or roasted on a barbecue are also present. Corpus entries include *grilled steak* (ĆULBASTIJA) [2] (SRP18980: 32 and SRP18992: 45), and very early on *beefsteak* (or *lungenbraten*)[15] (BIFTEK), as well as *pork chops* ("KARMENADLE") [1] (SRP18941: 88, 121), but not *schnitzel* (ŠNICLA) resulting from a special way in which meat is cut and prepared. Popović-Midžna tried to introduce the terms PRŽOLJICA and ROŠTILJAČA as Serbian equivalents of the German word *Schnitzel.* But she is inconsistent in their use, since they refer to other pieces of meat, as well. Thus, ROŠTILJAČA, for example could be either leg of pork/lamb or sirloin steak, while PRŽOLJICA is both leg of pork/lamb, JETRENICA (the term introduced by Popović-Midžina to mean *beefsteak*) and cutlet... The contemporary range of differently prepared chops was not part of the usual meal at the time.

The fat used to prepare food is animal fat [51] or butter [85], rarely kaymak. Animal fat is used for medicinal purposes and greasing; it is also necessary food in poor people's households (SRP19203: *150). In several examples, fat is used in the course of food preparation as in (SRP18881: 24), (SRP18941: 176) and (SRP19000: *172). Suet appears in stock phrases ŽIVETI KAO BUBREG U LOJU (literally *to live like a kidney in suet* i.e. *to live in the clover*) or IDE KAO PO LOJU (literally *(everything) is going (smoothly) as if greased with suet* i.e. *everything is going like clockwork*). When burnt for lighting, it is the material from which tallow candles (LOJANE SVEĆE or LOJANICE) are made, but as a source of fat it is not used when preparing food.

Additional thirty or so instances of the noun *chicken* (PILE) and adjective *chicken* (PILEĆI) in the dietary repertoire at the time are worth mentioning. To these, around fifteen instances of the words such as *wing*, *thigh*, *leg*, *rump*, *white meat* (KRILCE, BATAK, TRTICA, BELO MESO) should be added. An example from (SRP18934: *158) is particularly interesting

[15] Both Драгановић (1855) (Draganović, 1855) and Поповић-Мицина (1891) (Popović, 1891) use the term *lungenbraten*, meaning *beefsteak* in Austrian German. In (Поповић-Мицина 1891), however, it is not very clear whether it refers to beefsteak or a bigger piece of meat that incorporates beefsteak (for which the contemporary term LEĐANICA is used nowadays).

since it provides information that kaymak too, like butter in other national cuisines, could be used as fat when frying. As previously indicated, *frying in breadcrumbs* (POHOVANJE) is mentioned as a cooking method in (Поповић-Мицина 1891), but it is referred to as simply) *frying* (PRŽENJE). The only subsequent appearance of the term POHOVANJE in the corpus, significantly later compared to (Поповић-Мицина 1891) is found to be in the negative context in (SRP18941: 85) (The example from Section *Examples (Section 1)*).

Examples are given in Section *Grilled and Fried Meat Examples* in the Appendix.

Ajvar and Salads

Ajvar (AJVAR or HAJVAR) [6], which is nowadays considered to be a traditional national specialty does not appear in the corpus in its modern meaning. All recorded instances are related to the notion of *caviar*, as that meaning precedes the contemporary one (Zirojević 2020). Moreover, all instances happen to be in the context of three rich people's feasts (SRP18741: hajvar), (SRP19131: 41; SRP19190: 159: ajvar). In 19th century cookbooks too, this term is used exclusively in relation to caviar. Ajvar made of peppers is mentioned for the first time in (Поповић-Мицина 1911) where it is made using peppers and eggplants, with an unusual suggestion that fried eggplants should be peeled with a small silver spoon.

Salad (SALATA) [12] is another notion linked to middle class feasts, especially celery salad. Outside of that context, it appears in the corpus only once, as part of a lunch and twice as the garden crop, *lettuce* (ZELENA SALATA). *Sauerkraut*, *pickled cucumbers*, *pickled peppers* and *tarator* are also prepared as salads, but *tomato salad* or *Serbian salad* (SRPSKA SALATA) is absent from the corpus and cookbooks alike.

3.4 Alcoholic Beverages

Contrary to the contemporary view of *fruit brandy* (RAKIJA) as the national Serbian drink, the corpus provides a different picture. While there are 750 instances of *fruit brandy* (RAKIJA), *wine* (VINO) appears in it twice as often, 1298 times. This number of recorded instances of RAKIJA includes other names such as: ŠLJIVOVICA, PREPEČENICA, KOMOVICA, MUČENICA, LOZOVAČA and also MASTIKA and ANASONLIJA. Unlike RAKIJA, the noun VINO is often accompanied by an adjective determining the region of origin (Metohija → metohijsko, Negotin → negotinsko, Krajina → krajinsko, Župa

→ župsko, Primorje → primorsko, Bitolj → bitoljsko, Tokaj → tokajsko) or nominal determiners that still represent trademarks or brand names today (*Smederevka*, *Žilavka*, *Crvenika*, *Magyarater*, *Bermet*).

In several sources, the relationship between these two alcoholic beverages has a social dimension. Wine is drunk by the well-to-do, while brandy is for the poor (SRP18740: *141), (SRP18911:* 83), (SRP19071: *69).

Spritzer (ŠPRICER), i.e. wine mixed with soda/mineral water still did not have a name back then (SRP18911: *66), but both local and imported brands of mineral water were used to prepare it.

The surprising thing is that, besides wine, another drink that was frequently enjoyed was *champagne* (ŠAMPANJAC [35]: ŠAMPANJ, ŠAMPANJER, ŠAMPANJSKO VINO, PENUŠAVO VINO). Its frequency is noticeable in the first part of the corpus from (SRP18520) to (SRP18911) and then it vanishes until (SRP19090) when it appears again, remaining present until the end of the time period covered by the corpus in (SRP19201: 714) (see Figure 2).

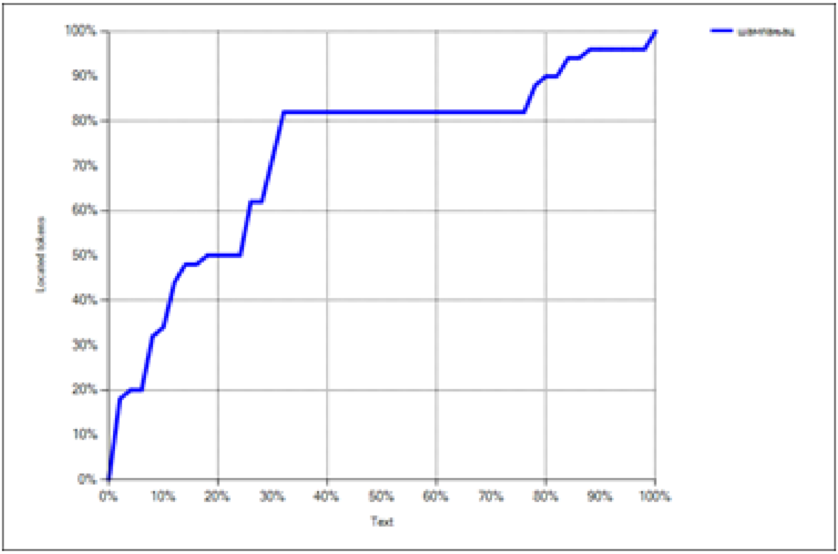


Figure 2. Cumulative frequency for the filter <<`šampanj`>>.

Other foreign alcoholic beverages were not unknown. Thus, *cognac* [7], *rum and grog* [11], as well as *whiskey*, *cherry brandy*, together with *punch*, *curaçao* and *absinthe* are also present.

Examples are given in Section *Alcoholic Beverages Examples* in the Appendix.

4 Conclusion

This short overview of the food and beverages attested in the SrpELTeC corpus bears witness to what was recognized as a nutrition-related concept in Serbian at the time. Moreover, what is dealt with here is not individual knowledge, but concepts that writers can share with their readers. On the other hand, the corpus clearly delineates the zones of cultural influences originating from Turkey and Austria-Hungary. Unlike early ethnographic writings, where we find excitingly simple descriptions of dietary choices, these corpus attestations illustrate their social (and lexical) complexity to the fullest.

The analysis of the SrpELTeC corpus provides additional information that helps paint a more complete picture of the dietary habits of the Serbian population from 1850 to 1920 as described in different ethnographic and historical sources, taking into account not only the choice of foodstuffs and cooked meals, but also their taste. Even at the level of the processed material, it is possible to perceive what is usual, new or peculiar, as far as eating habits and modernization of food preparation methods are concerned. Besides, an element missing in other sources, namely, nutrition of urban population is amply illustrated, including its social dimension.

Similarly, this corpus can surely render the picture of other aspects of life in the second half of the 19th and early 20th century more complete. Complex issues, such as the position of women in the Serbian society, children's education, cultural habits, medical treatments, means of travel and many other questions are amply illustrated too, opening possibilities for conducting further analyses of the corpus.

References

Dikas, Alen. 2016. *Rečnik zaljubljenika u kulinarstvo.* Beograd : Službeni glasnik.

Flandrin, Jean-Louis. 2002. *L'Ordre des mets.* Paris: Odile Jacob.

Krstev, Cvetana. 2021. “White as Snow, Black as Night – Similes in Old Serbian Literary Texts.” *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.6.

Krstev, Cvetana, Duško Vitas, Miloš Utvic, and Branislava Šandrih. 2017. “The New Clothes for an Old Cookbook.” In *Proceedings of 8th Language & Technology Conference, November 17-19, 2017, Poznań, Poland,* 174–178. http://poincare.matf.bg.ac.rs/~cvetana/biblio/ltc-042-krstev.pdf.

Mijo, Kristijan. 2012. *Rečnik zaljubljenika u gastronomiju.* Beograd : Službeni glasnik.

Montanari, Massimo. 2011. *Hrana kao kultura.* Zagreb : Sandorf.

Vitas, Duško, and Cvetana Krstev. 2012. “Processing of Corpora of Serbian Using Electronic Dictionaries.” *Prace Filologiczne* XVIII:279–292.

Zirojević, Olga. 2019. *Istočno-zapadna sofra: mali kulturno-istorijski i kulinarski leksikon.* 2. izd. Beograd : Geopoetika izdavaštvo. ISBN: 978-86-6145-290-1.

Zirojević, Olga. 2020. *Od sumnjivog napitka do kavijara: iz osmanske baštine.* Zemun : Udruženje za kulturu povezivanja Most Art Jugoslavija. ISBN: 978-86-80640-39-6.

Белић, Александар et al., ed. 1959–2021. *Речник српскохрватског књижевног и народног језика.* Београд : САНУ, Институт за српскохрватски језик. ISBN: 86-7025-309-7.

Витас, Душко М. 2018. “Храна из нежељене поште : (анатомија језика брзе хране).” In *Српски језик и његови ресурси: теорија, опис и примене,* edited by Божо Ћорић and Александар Милановић, 47:21–35. Научни састанак слависта у Вукове дане 3. Београд: Међународни славистички центар, Филолошки факултет, Универзитет у Београду. ISBN: 978-86-6153-521-5. https://doi.org/10.18485/msc.2018.47.3.ch2.

Грђић-Бјелокосић, Лука. 1908. “Српска народна јела у Херцеговини и у Босни.” In *Српска народна јела и пића. Књ. 1,* edited by Јован Ердељановић, vol. 10. Београд : Српска краљевска академија.

Драгановић, Јеротеј. 1855. *Србскій куварЪ: (по немачкому кох-бухЪ).* У Новом Саду : ТрошкомЪ Ігнаца Фухса, кнЬижара.

Зиројевић, Олга. 2005. “Јело и пиће.” In *Приватни живот у српским земљама у освит модерног доба,* edited by Александар Фотић. Београд : Clio.

Карић, Владимир. 1887. *Србија: опис земље, народа и државе.* Београд : Краљевско-српска државна штампарија.

Костић, Ђорђе С. 2019. *Трпеза за уморне путнике : европски путописци о исхрани у Србији у 19. веку.* Београд : Геопоетика издаваштво.

Крстев, Цветана, and Биљана Лазић. 2015. “Глаголи у кухињи и за столом.” *Научни састанак слависта у Вукове дане–Српски језик и његови ресурси: теорија, описи, примене* 44:3:117–135.

Марковић, Спасенија Пата. 1959. *Велики народни кувар [Патин кувар].* Београд : Народна књига.

Мијатовић, Станоје. 1908. “Српска народна јела (са прилогом о пићима) у Левчу и Темнићу.” In *Српска народна јела и пића. Књ. 1,* edited by Јован Ердељановић, vol. 10. Београд : Српска краљевска академија.

Мирковић, Софија. 1922. *Велики српски кувар.* 3. поправљено и проширено изд. Нови Сад : Издавачка књижарница С. Ф. Огњановића.

Петровић, Петар М. 1937. *Свезнање: општи енциклопедиски лексикон у једној књизи.* Београд : Народно дело.

Поповић-Мицина, Катарина. 1891. *Велики српски кувар : са млого лепих и врло вештачки израђених слика : за употребу српских домаћица.* Нови Сад : Браћа М. Поповић.

Поповић-Мицина, Катарина. 1911. *Велики српски кувар : са много лепих и врло вештачки израђених слика : српским домаћицама : са допунама најновијег практичног кувања.* Нови Сад : Учитељско деоничарско друштво „Натошевић“.

Тројановић, Сима. 1983. *Старинска српска јела и пића.* Vol. 3. Београд : Просвета.

Фотић, Александар, ed. 2005. *Приватни живот у српским земљама у освит модерног доба.* Београд : Clio.

Appendix

Examples (Section 1)

(SRP18935: 86) „О, да ми је сад комадић окореле проје, парче сира, сланине, ја сува меса, те бих се сита најела !...“

(SRP18941: 85) Поп Спира није марио за *поковано пилеће*; као Србин и православне цркве син мрзео је на то *швапско печење.*

(SRP19201: 140-1) [...] онде по неко усамљен, ослоњен на зид, на дирек општинскога фењера, ломи рукама хлеб који држи под мишком и штрпка тврди сир из жуте бакалске хартије. А тај исти је, само неколико дана раније, са белом сервијетом на коленима, немарно прелетао очима јеловник, враћао порције што му је дато мршаво месо и три пут враћао келнера да му донесе хлеб са меком горњом кором. [...] Једна београдска породица, која се виђа о премијерама у позоришту, у чијим салонима господаре Шопен, Григ и Чајковски, која отказује службу девојци што је слаткише за жур купила у тој и тој радњи, где никад нису довољно свежи — очајно је и сатима дреждала пред ћепенком једне ћевабџинице, не би ли дошла на ред да купи неколико козјих ћевапчета, које ће из једних старих новина, на сред приштинске улице, јести прстима.

(SRP18960: 76-7) Учитељ пропитиваше из предавања о овци.

— На колико се делова дели овца?

— Овца се, молим господине, дели на пет делова на: главу, врат, труп, ноге и реп... [...]

— Децо! Једе ли се во ?

— Во се једе — рече један ђак, па и остали

— Једе ли се медвед?

— Не једе — накељише се деца.

— А, на пример, магарац?

(SRP19100: 203) Вино је било нешто накисело, сир тврд, а кајмак преслан. Ипак се њима све то чинило врло добро. Милош је говорио да тако питког вина нигда није пио, Зорка је тврдила да се тако добар сир не може наћи на пијаци, а обадвоје су се слагали да је кајмак одличан.

Examples of Food of Plant Origin

<1> (SRP19102: 313) Тако се и у старој градини направило више места за цвеће, *поврће*, *варива*, малине итд.

<2> (SRP19102: 50) И воће, и *зелен*, и цвеће — све је најбоље врсте и марљиво гледано и неговано.

<3> (SRP19012: 3) Недеља је, пост, говорило се о посној чорби, о ћуфтетима од сецкана *празилука*, о *црноме луку* на тепсичету [...]

<4> (SRP19012: 99) И свекрва ми од собајле у башчу: садила *арпаџик* и *бели лук*, сејала *ћеревиз* и *магданос* [...]

<5> (SRP18891: 120) — Ово је моје цвеће! — рече попа. Ово се цвеће зове *купус*, ово *кромпир*, ово *пасуљ*, или како га ми сељаци зовемо *гра*, ово *першун*, ово, што лепо мирише, *мирођија*, ово *ротква*, ово...

<6> (SRP19100: 167) [...] баште по којима расте *боранија* и *празилук* [...]

<7> (SRP19130: 86) У свима је, и онда као и сад, био најпростији намештај сеоских кућа: неколико лонаца и чинија, наћве, ковчег за брашно, вреће с *пасуљем* и *сочивом*, низови *лука црног* и *белог* [...]

<8> (SRP19140: 68) Ниже од њих растао је *грашак*, са широким зеленим *махунама*, које су висиле на све стране као фантастичне ресе.

<9> (SRP18880: 94) Ја је замењујем, прибирам *боб* и *пасуљ*, сушим *мехуне*, млатим их, и зрна остављам за зимицу.

<10> (SRP19101: 163) И предаде јој сито, пуно шећера, *леблебија* и осталих шећерлема [...]

<11> (SRP18992: 30) Трешња у цвету, *купус* тек посађен и примљен.

<12> (SRP18950: 6) Између осталих прочитао је и то: како се могу очувати *зелене паприке*, па да буду свеже усред зиме, као да су тога часа у башти узабрате, [...]

<13> (SRP19203: 776) [...] поред тога знам да је Бугарин превртљив, па хоће и да превари и да изневери. Познајем ја њих, били су неки и код нас, што саде *патлиџан* и *паприку*.

<14> (SRP119193: 94) Ја је замењујем, прибирам *боб* и *пасуљ*, сушим *мехуне*, млатим их, и зрна остављам за зимицу. Да видиш моје лехе, моје *патлиџане*, моје *лукове*, мој *кромпир* — све кућице у правилним редовима[...] Поред плота засадила сам *сунцокрет*, већ је израстао више моје главе, још који дан, па ће своје златне круне повијати према сунцу.

<15> (SRP19090: 145) ... двоје дјеце ваде запретане *крумпијере* из жераве...

<16> (SRP18932: 142) Анђелија прво оде у башту те нађе *краставац* и запрета га у пепео....

<17> (SRP19012: 62) - Те си напраји' један таратор за ћеф од зелене [...]

<18> (SRP19041: ?) Вртлићи, што би их зец прескочио, засађени *зељем*, леже неуређено и расијано у осјену кућа.

<19> (SRP19203: 44) [...] сејао је *ротквице*.

<20> (SRP18730: 56) [...] на њој је било да изда све наредбе о врту [...] где су сејани купус, краставци, диње и лубенице, роткве и остало.

<21> (SRP18741: 118) *Карфиола* и *шпаргла* мора имати, па ма шта коштало [...]

<22> (SRP19001: 29) Брзо одрешим шерпу у којој беше сав онај *кељ* што сам га видео и пржено месо.

<23> (SRP19140: 279) [...] они су ишли дуж некога зида, преко кога су се спуштале процветале вреже од *тикава*.

<24> (SRP19071: 9) [...] примаче преда се бедну своју вечеру, комад хлеба и повеће парче печене *бундеве* [...]

<25> (SRP18960: 13) На столу, до прозора, чаша воде с цвећем, а око ње дуње и *шућурке* [...]

<26> (SRP19193: 6) Код јабуке један вечити џбун крупног сјајног високог лишћа од *рена* даје од монотоније уморном оку, у сва годишња времена лепе зелене боје у свима нијансама редом.

<27> (SRP18760: 313) Код старе госпође био је ручак већ спреман, супа готова, а у једној великој кастроли је *спанаћ* [...]

<28> (SRP19100: 129) [...] у прашини сунчевих зрака, жутила се ниска поља, засађена *салатом* и *купусом*.

<29> (SRP18936: 50) хтеде је бесно погледати али му очи, одоше на салату од *целера*.

<30> (SRP18880: 72) — Па ђувече од *црвених и модрих патлиџана*? Ево како се то готови... — Али докле ћеш ме, Саво, бити по ушима са тим твојим *модрим патлиџанима*? Увек си био онако, онако.... [...] Што ти се сад наспеле те луде цинцарске сплачине?

<31> (SRP18941: 205) — Та задрж'о ме овај проклети *парадајз* код куће, и баш сам данас нашла да га кувам!

<32> (SRP19012: 26) [...] бре прајила сам слатко; бре варила сам *црвени френци*; [...]

<33> (SRP18870: 144) Масло је врло добро и свеже. Бојадишу га *жутом репом* и *сафраном*, ал ја то осетим по мирису, па волим платити новчић два више, само да је добро и чисто.

<34> (SRP18760: 374) Него, ако немаш георгине, а ти узми црвеног *божура* и два-три струка *сунцокрета*.

<35> (SRP19140: 69) Око плотова су расли *сунцокрети*.

<36> (SRP18992: 14) [...] једите црн сухи хлеб *овсени*, *ељдани* или *никакви*.

<37> (SRP18964: 149) И *просенице*, и *сира* и *лука* и *соли* и вина!

Examples of the Simplest Meals - Bread and Onions

(SRP18935: 40) За тим домаши своју торбу, завуче руку унутра, извади комадић *паучљиве проје*, *једну главицу црног лука*, одреши крпицу у којој је *со* завезана била, и поче авољити.

(SRP18971: ?) Тако се отимала од немаштине и глади, а дао Бог — сеоској души не треба много: *комад проје* и *главица лука* задовољава потпуно њене потребе и навике. Чутура иде редом, неки присмаче и *хлеба* и *лука* те залива ручак црним вином.

(SRP19001: 192) Два пуна месеца радио сам са зидарима и живео о *хлебу* и *луку* као и они

(SRP19131: 122) Да завиди својим надничарима кад их гледа како слатко једу *сува хлеба* и *лука*, јер он, кад седне за своју милионарску трпезу не може ни да окуси ништа од најскупљих ђаконија, јер није никада гладан.

(SRP19132: 297) Ни ручавати није хтио за софром, с осталом чељађу, него носио у торби, о рамену, *овећи комад хљеба* и *неколике главице лука*, па ручавао негдје осамљен, сакривен од свакога.

Examples of the Simplest Meals - Bread, Onions and a Little Something More

(SRP18881: 83) Бележниковица га задржала на *сир с црним луком*, да прави бележнику друштво.

(SRP18920: 172) Доле се посадише на шареном сиџадету око једног белог убруса, на који Рајко положио беше *нешто погаче*, *млада црна лука*, *соли* и *млада сира*.

(SRP18940: 100) Беше ту у једном дрвеном чанку, лепа као *кајмак*, *папула од белог гра*. Неколико *главица бела и мрког лука*. У супрету *испечених и укуваних кромпира*, и као снег бела *погача*. У једној чистој кринци беше завијено *соли*.

(SRP19080: 182) На сваком столу стоје тепсије с питом, гужваром, колачима, а на свему томе *куван кромпир и лук*, а крај тепсије у тањиру *папула* — народна храна. Тим се сељак храни на дому па с њиме и славу слави.

A Description of a Feast

(SRP18911:65) Да се не би учинило на жао каквом гурману, изређаћемо сва јела, која су донесена после киселе чорбе, на овој вечери.
Пилећи паприкаш са резанцима предњачио је, за њим дође винова сарма, једна реткост у ово доба, но вешта домаћица имала је начина, да овај зимњи шпециалитет одржи. Коме да вода на уста не поцури, кад се спомене, да се за овим појавила пита с месом. Ово српско исторично јело, својим мирисом и својом унутрашњом структуром управо очарава све живце, а задатак му је, да помоћу својих чаробних кључева отвара даљи апетит. Сва је дворана замирисала, кад се ово благородно јело унело, да је сирото „реш“ печено прасе, које је за овим дошло, имало читаву муку, да буде поједено. Јагњеће печење, опет једна реткост у овом добу године, нашло је одзива само код старих господара и госпођа, јер је омладина била већ толико испуњена одушевљењем, питом и прасетом, да је само још чекала, да музика засвира, па да се отпочне набијање унесених количина, и наступи прво варење црним неготинским вином знатно потпомогнуто.
Са свим меланхолично, и без икаквог утицаја на сите госте, лежаху по столу поређане „париске штанглице“, „патишпан“, „торта с орасима“, „пуслице“ и „ванили ринглице“.

Examples of Food Staples or Groceries

(SRP19102: 382) Посуђе је сад мало другојаче и много се чистије држи, а стоји по полицама, које се често перу. И јела се више не кувају нити онако масна, ни слана, ни љута, а готове се много боље и посна и мрсна. Нарочито више пазе, барем ученице Даничине, да сваког дана буду друга јела и да она буду увек свежа. Оне сад умеју и да месе сем гибанице и друге колаче, особито с медом и воћем.

(SRP19193: 20) Стара сам. Научила сам на *кафу* и на *ракију* сваки дан, а и да поједем лепо, и да се обучем топло.

Examples of Dishes

(SRP18740: 133) У бакрачету је већ врела вода. Стојанова мајка захвати из обешене јареће мешине две-три прегрши пројина брашна и сасу их у врелу воду. Стојан узе мирно једно парче чисто остругана дрвета, промеша својим снажним рукама, скиде са верига бакраче и изручи качамак на совру, на којој већ беше постављено нешто мало сува меса, обарених јаја, лука, паприке и соли... И тако та мала породица седе за совру, да вечера:

(SRP18967: ?) Нигде се у свем свету толико не поједе као код нас. Сва јела, која год познајемо, скувамо на једаред, само ако имамо. Ето“, каже, „славиће наш ратар свечарство или има у кући сватове, па одмах зову куварицу из вароши и почну износити јела у подне, па буде ноћ, а још се није све изређало. Наше новине“, каже, „не доносе шта ручају цареви и краљеви кад се састају и госте, а требало би да доносе. Не што се то нас тиче, него да види свет, да се ни при царским гозбама не износи више од четири јела; за тим долази сир и воће. Ми ни на што друго не дајемо, до ли на јело. Најбоље, најздравије собе наше, те су намештене, у које и не улазимо, а живимо у вајатима и коморама. Чак и наши богатуни, који имају по три-четири намештене и гостинске собе, немају у кући купатила, немају књижнице — ништа.“

Soup and Broth Examples

(SRP18741: 121) *Црна чорба* са рибаним пармезаном не сме фалити; па онда гарнирана говеђина, красна гарнитура!

(SRP18760: 377) баш као она масна ђинђувица на *дебелој говеђој чорби* [...]

(SRP18871: 41) Имали смо најпре *чорбу јагњећу*;

(SRP18940: 54) „Што је више јаја, гушћа је *чорба*,“

(SRP18964: 50) [...] бејаше *чорба од свињског меса* а у другој од *рибе.*

(SRP18960: 116) Кад немате *супе*, онда иде *чорба*: или *пилећа*, ил’ *гушчија*, *јагњећа*, *ћурећа*...

(SRP19012: 3) говорило се о *посној чорби* [...]

(SRP19102: 65) За вечеру најпре принеше као обично сира с младим кајмаком, па онда *чорбу пилећу зачињену кајмаком и јајима* [...]

(SRP18950: 34) Ама у моје време није се знало за беле кафе и крофне, нити се знало за те милипроте, ни за те *супе* и *сосове*!

(SRP18961: 26) Наш је народ у граници, бар у сремској, уобичајио, да у оваким приликама долази прво *чорба*, тако звана „*супа*“ на сто.

(SRP19001: 27) Једног дана [...] добијем за ручак само ону *зелен из супе* са оним талогом што остане на дну лонца кад се *супа* оцеди.

(SRP19140: 269) питала своје гошће шта више воле: *киселу чорбу* или *супу с кнедлама* [...]

Grilled and Fried Meat Examples

Kebab **(ĆEVAP) and** ***Grilled Minced-Meat Finger*** **(ĆEVAPČIĆ)**

(SRP18690: 5) мени је [...] добро као да сам појео педесет *ћевапчића*

(SRP18882: 37) на пољу ће се пећи свињски *ћевап* и *печење од живине* на ражњу

(SRP18882: 38) Гледа како ђутуричар креше ражањ па надева *ћевап*. Како које парче надене а он га поспе мало ситним луком, зеленом паприком и закити парченцетом сланине. (SRP18934: 88) А, ми смо смазали два ћевапа. У сваком је било по аршин [...]

(SRP18932: 25) у једној руци држи *ћевап* а у другој руци чутуру. Сватови стадоше. Неки узеше по *режањ од ћевапа* и — чутура пође од руке до руке.

(SRP18940: 170) [...] сваки држи по оканицу вина и ракије и по врућ *ћевап на ражњу*...

(SRP19140: 269) [...] те је питала своје гошће шта више воле: [...] *ћевап у дунсту* или *ћурче на подварку.*

Roasting and frying

(SRP18881: 24) [...] цврче *батаци* на угрејаној масти,

(SRP18934: 158) Могло би се, на брзу руку, *попригати* које *пиле* на *младом кајмаку*

(SRP19000: 172) [...] долећу често они сеоски »колачићи« умешени с јајима, испржени на масти, што их господа варошка зову уштипцима...

(SRP19000: 172) Кроз прозорче се једнако провлаче дугачки судови са читавим »кршем« *разноврсна печења*: *прасећег*, *пилећег*, *телећег*... па се ту жуте и преливају са руменом, масном корицом добро *ућљукане ћурке* и *гуске*, па неке *младе дебеле пловчице* пржене на кајмаку [...]

(SRP19203: 150) Лепо сам пазарио, и кући сам донео килу *соли*, *пун лонац масти*, џакуљицу брашна [...]. Сиромах сам човек, али ми је драго, кад у кући има брашна, масти и соли.

Alcoholic Beverages Examples

(SRP18740: 141) — Хвала, господине, а баш ти слабо марим за вино! Вино је за господу; а ја ти волим чашицу ракице, него целу вучију вина.

(SRP18911: 66) Старци пију црно вино само кад су слаби, па им то лекар пропише, а иначе су већином за бело, па му неки додају још и минералне или обичне воде; тако н. пр. Недељковић меша с вином чисту воду, Богатић пије уз бело вино „гисхиблер“ а стари начелник и домаћин „Буковичку воду“.

(SRP18911: 83) У локалу се пије понајвише ракија, сиротињско пиће.

(SRP19071: 69) [...] пиво је скупље, а вино јевтиније, и, што је главније, не дангуби — добива у времену, јер се од жупског вина много раније опије него од оног глупог швапског пива. И сада се лепо виде на њему обе те периоде пића: од пива је сачувао трбух, а од жупског вина стекао нос црвене, као бакар, боје.

Scientific paper

White as Snow, Black as Night – Similes in Old Serbian Literary Texts

UDC 811.163.41’322.2
DOI 10.18485/infotheca.2021.21.2.6

Cvetana Krstev
cvetana@matf.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

ABSTRACT: This paper outlines the use of simile rhetorical figure in the Serbian subcollection SrpELTeC, a part of the ELTeC collection. We analyze their use by different criteria: authors, time periods, author’s gender, novel’s size and novel’s popularity. We also analyze adjectives, nouns and markers used in similes found in SrpELTeC, as well as entities to which they apply. We briefly compare this results on an *ad hoc* sample of contemporary Serbian novels.

KEYWORDS: literary corpora, simile, local grammars, ELTeC, Serbian language

PAPER SUBMITTED: 22 November 2021
PAPER ACCEPTED: 28 December 2021

1 About Simile

Veale and Hao (2008, 253) describe similes as “a window to the folk knowledge, since explicit similes make use of highly evocative and inference-rich concepts to ground comparisons and make unfamiliar seem familiar.” They add that “the simile [...] is one common vehicle for folk wisdom, one that uses explicit syntactic means” (454). To illustrate comparison as rhetorical figure in Serbian language, and corroborate these claims, we will give an exquisite example: *U njenim očima on je bio: visok kao bor, mio kao proleće, dobar kao Anđeo hranitelj, mlad kao rujna zora, beo kao labud, lep kao prolećni dan, hrabar kao Obilić!* (In her eyes he was: tall as a pine, dear as spring, good as Angel fosterer, young as ruddy dawn, white as a swan, beautiful as a spring day, brave as Obilić!).[1]

[1] This sentence using 7 similes in sequence is from the novel *Vojnik Stojan: nedovršen ratni roman* (Soldier Stojan: an unfinished war novel) by Dragomir Petrović (1918), which was not used in this research.

In (Krstev, Jaćimović, and Vitas 2020) we presented a preliminary research on the use of simile rhetorical figure in the incomplete version of the SrpELTeC sub-collection,[2] which contained 41 novels. In this paper we will repeat and enhance this research on the 100 novels SrpELTeC sub-collection.

As pointed in (Israel, Harding, and Tobin 2004) both literal comparison and simile have the same recognizable formal structure, a surface form consisting of the following elements: the subject of comparison (TARGET, TOPIC or TENOR), the object of comparison (VEHICLE or SOURCE), a conjunction which signals a comparison (MARKER, in Serbian usually *kao* (as)), and the basis of the comparison implied by the expression (PROPERTY or GROUND), as illustrated by the following example:[3]

reče	*Pavle*	*beo*	*kao*	zid
said	Pavle	white	as	wall
	TARGET	GROUND	MARKER	VEHICLE

As pointed in (Brehmer 2009), the TARGET is usually not a part of a simile (we will corroborate this in Subsection 3.2). Similes usually have a closed structure, containing all three elements: GROUND, MARKER, VEHICLE, as in the previous example, but could also be open, if the attribute is not explicitly stated, but could be derived from the context (MARKER, VEHICLE), as in the following example:

Arhimandrit	*beše*	*(ljut)*	*kao*	*ris*
The archimandrite	was	(angry)	as	a lynx
TARGET		missing GROUND	MARKER	VEHICLE

In this paper we will consider only closed similes.

Besides adjective similes, which represent a class of multi-word expressions (MWE), verbal multi-word expressions (VMWE) are also used for comparison, forming simile figures if conventionalized, as pointed in (Niculae and Yaneva 2013) for English, (Mitrović, Markantonatou, and Krstev 2020) for Greek and (Мршевић-Радовић 1987) for Serbian. In this paper we will not deal with this type of similes, although we will briefly compare the two types at the end of Section 3.

[2] ELTeC is a multilingual collection of novels published in the period from 1840 to 1910. It is developed in the scope of the COST action CA 16204 *Distant Reading for the European Literary History*. SrpELTeC is the sub-collection containing novels in Serbian. More about this sub-collection can be found in (Trtovac, Milnović, and Krstev 2021) in the same issue.

[3] All examples in this paper will be from the SrpELTeC.

This paper is organized as follows. In Section 2 we will present the setting of our research and the methods used. How similes are used in SrpELTeC will be discussed in Subsection 3.1, while characteristics of these similes will be analyzed in Subsection 3.2. A brief comparison of the use of similes in Serbian novels from 19th to early 20th century and novels from the second half of the 20th and the beginning of the 21st century will be given in Section 4. Some directions for future work will be mentioned in Section 5.

2 Research Methods

ground	**marker**	**modification**	**vehicle**	**modification**
hitar	*kao*		*jelen*	
hitar	*kao*	*mlad*	*jelen*	
fast	as	(a young)	deer	
slobodan	*kao*		*ptica*	
slobodan	*kao*		*ptica*	*u gori*
free	as		a bird	(in a wood)
beo	*kao*		*sneg*	
beo	*kao*	*najbelji*	*sneg*	*u planini*
white	as	(the whitest)	snow	(in a mountain)

Table 1. Modifications of a vehicle in similes

The basic structure of simile figures – GROUND, MARKER, VEHICLE – can sometimes be modified, and modification concerns mostly the vehicle (or source). Two most frequent types of simile modifications are represented in Table 1.

Similes can also occur in variants, which can be the result of different pronunciation (Ekavian or Ijkevaian), use of dialects or variant forms, diminutives, etc. Some cases of variants are represented in Table 2.

In addition to that, similes do not always appear in a text in the expected word order (adjective – conjunction – noun), for instance:

. . .	*kao*	*sneg*	*belih*	*grudi*
. . .	as	snow	white	bossom
	MARKER	VEHICLE	GROUND	TARGET

Also, the main components of similes are not always contiguous, since insertions are possible:

ground	**marker**	**source**	**type**
beo	*kao*	*sneg*	Ekavian
bijel	*kao*	*snijeg*	Ijkevaian
white	as	snow	
hladan	*kao*	*led*	
ladan	*kao*	*led*	variant (non-literal)
cold	as	ice	
mlad	*kao*	*kap*	
mlad	*kao*	*kaplja*	synonym
mlad	*kao*	*kapljica*	diminutive
young	as	a drop	
beo	*kao*	*zid*	
beo	*kao*	*duvar*	synonym
white	as	wall	

Table 2. Variations of similes

...	*a*	*tanak*	*je*	*kao*	*prut*
...	and	slander	is	as	a twig
		GROUND		MARKER	VEHICLE

Some similes are not complete because occasionally two similes are contracted when they use the same vehicle:

brzo	*i*	*vešto*	*kao*	*mačka*
fast	and	skillful	as	a cat
GROUND		GROUND	MARKER	VEHICLE

Having all these in mind we used two methods to retrieve similes from SrpELTeC:

- We used local grammars in the form of finite-state automata implemented in Unitex,[4] which take care about all modifications, variations and possible changes in the text mentioned before and which were developed earlier on the smaller sample of novels (Krstev, Jaćimović, and Vitas 2020). The set developed within mentioned research comprised of 243 local grammars for the recognition of similes in Serbian texts.
- we used simple patterns to retrieve other possible occurrences of similes. These patterns were implemented in Unitex as well, and they rely on the

[4] Cross-platform corpus processing suite Unitex/Gramlab.

Serbian electronic dictionaries (Krstev 2008). One such simple pattern is:

```
<A>
(<jesam.V>+<E>)
(kao+ko+k('+')o+ka('+ka')+ka+nalik+poput+kano)
```

The pattern is composed of an adjective followed by the conjunction *kao* (in various forms) or other prepositions, with an optional form of the auxiliary *jesam* (to be) in between. The obtained results were manually filtered to reject false recognition.

3 Analysis of Results

3.1 Distribution of Similes in SrpELTeC

Using methods described in the previous section we retrieved 1,051 occurrences of the simile rhetoric figure, an average of 10.5 occurrences per novel. In five novels no simile figures were found: *Jedna ženidba* (SRP18620), *Jurmusa i Fatima* (SRP18790), *Srbin i Hrvatica* (SRP18921), *Pokojnikova žena* (SRP19022), *Stradija* (SRP19025). The highest number of occurrences – 59 – were retrieved from the novel *Hajduk Stanko* (SRP18963), followed by *Nove* (SRP19120) – 39 – and *Novac* (SRP19060) – 37. When sorted by relative frequencies,[5] the novel *Hajduk Stanko* still remains on the top – 5.98 – followed by *Radetića Mara* (SRP18940) – 5.75 – and *Borci* (SRP18891) – 5.48. It is interesting to note that the first and the third novel are written by the same author, Janko Veselinović, and his third novel in this corpus *Seljanka* (SRP18932) is ranked as 16th with the relative frequency 4.32. There are 29 novels with less than one simile per 10,000 words, and the relative frequency of similes in the whole collection is 2.20.

The relative frequency of simile in SrpELTeC novels per four corpus composition criteria[6] is represented in Figure 1. One can observe that the authors of T2 novels used less simile figures than the authors in other periods – data for T1 period is not really comparable since there are only two novels in that

[5] The relative frequency is calculated as the number of similes per 10,000 words.

[6] For the composition criteria of ELTeC collection and distribution of novels from SrpELTeC according to them see (Trtovac, Milnović, and Krstev 2021) in this issue.

period. Female authors tend to use fewer simile figures than male authors – data for authors of unknown gender is not significant since there is only one such author. It seems that longer novels use less simile figures, but it is hard to find an explanation for the dependency of the number of similes on a novel's size. One has to bear in mind that there are only 5 long novels in the whole corpus. It seems that the binary reprint parameter, representing the popularity and presumably the quality of novels, is not correlated with the use simile figures.

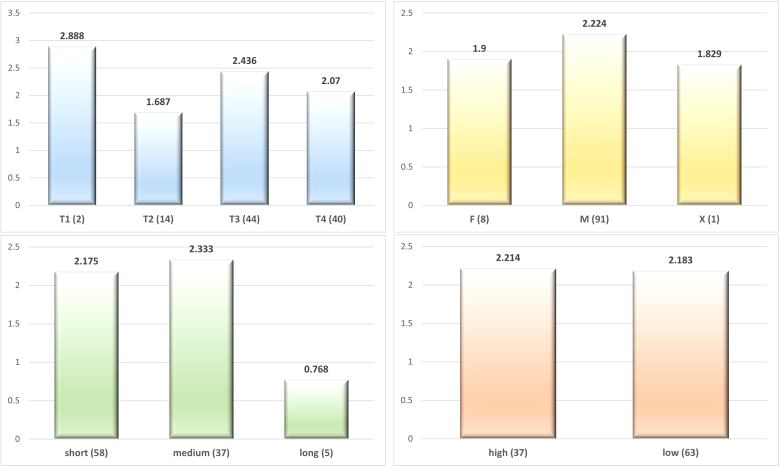


Figure 1. The relative frequency of simile in SrpELTeC novels per four corpus balance criteria: time slot, gender, size, reprint

When the novels are ordered by the year of publication and grouped into groups of 10, the average numbers of similes in groups appears to be uncorrelated with the year of publication (Figure 2, left). When the same is done for the size of novels, measured by number of words,[7] the result is the same – no correlation (Figure 2, right).

We ranked the authors who are represented in SrpELTeC by at least 3 novels according to the relative frequency of simile occurrences in all their novels, and the results can be seen in Figure 3.

[7] All novels are sorted by their size, and then grouped in groups of 10 according to their rank in the sorted list.

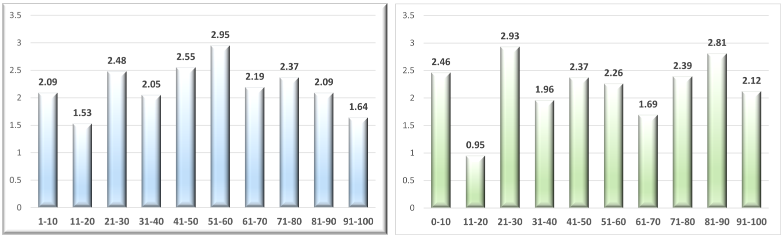


Figure 2. The relative frequency of simile in groups of 10 per: (a) year of the first publication; (b) number of words.

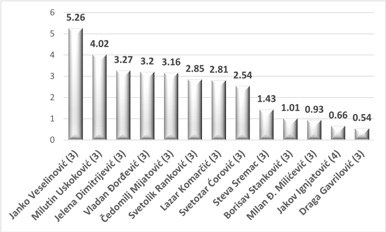


Figure 3. The relative frequency of simile for authors that have more than 3 novels in SrpELTeC.

As stated before, we retrieved 1,051 similes from SrpELtEC, of which 556 were different. We treated as equal similes that have the same property and the vehicle but can differ in the marker or due to modifications (see Table 1). We also treated as equal similes with different property and/or vehicle if that difference comes from different pronunciation or variant form (see first two rows in Table 2). There were 426 similes that occurred only once. Top ten similes in SrpELTeC are presented in Table 3. If we take the number of novels in which a simile appears as a sign of its popularity, then the obtained results show that the top 10 most frequent similes are also the most popular, although their rank is changed slightly, as seen in Table 3.

3.2 Characteristics of Similes

Ground – Adjectives A total of 202 different adjectives appear as properties among all extracted similes, with 102 of them in only one simile. The most frequent adjectives and nouns with which they combine and occur

simile	**translation**	**absolute frequency**	**popularity in novels**	**rank by popularity**
beo kao sneg	white as snow	48	30	1
bled kao krpa	pale as a cloth	37	22	2
bled kao smrt	pale as death	37	14	4
hladan kao led	cold as ice	23	15	3
crven kao krv	red as blood	14	12	5
jasan kao dan	clear as day	13	9	8
mlad kao kaplja	young as a drop	13	11	6
plav kao nebo	blue as sky	12	9	9
crven kao rak	red as a crab	11	8	10
ljut kao ris	angry as a lynx	11	11	7

Table 3. The most frequent and the most popular similes

more than once are presented in Table 4. It can be seen that all most frequent adjectives combine with a number of nouns (column **No.** in Table 4 displays the number of nouns with which an adjective combines), and that the most frequently used noun barely exceeds 50% of all occurrences (the number in column **%** in Table 4 represents the percentage of appearances of the most frequent noun among all occurrences). Moreover, similes in which an adjective combines with only one noun never occur more than twice.

adj.	**freq.**	**No.**	**%**	**nouns**
bled	103	16	35.9	37: krpa, smrt, 7: vosak, 5: mrtvac, 3: kip, 2: ljiljan, samrtnik, senka
pale				37: cloth, death, 7: wax, 5: dead person, 3: statue, 2: lily, dying person, shadow
beo	83	21	57.8	48: sneg, 9: mleko, 3: alabaster, ovca, zid, 2: krin
white				48: snow, 9: milk, 3: alabaster, sheep, wall, 2: lily
crn	52	22	13.5	7: gar, 6: noć, 5: ugljen, zift, 4: gavran, zemlja, 3: gak, trnjina, 2: ugalj
black				7: soot, 6: night, 5: coal, tar, 4: raven, earth, 3: grey heron, blackthorn, 2: coal
hladan	43	15	55.8	24: led, 5: stena, 2: grob
cold				24: ice, 5: rock, 2: grave
crven	37	10	37.8	14: krv, 11: rak, 4: paprika, 2:vatra
red				14: blood, 11: crab, 4: paprika, 2: fire

Table 4. The most frequent adjectives and nouns with which they combine

Similes are often used to describe colors. There are 9 adjectives representing colors in our corpus: *beo* (with Ijekavian variant *bijel*) (white), *crn* (black), *crven* (red), *plav* (blue), *zelen* (green), *žut* (yellow), *siv* (gray), *modar* (livid), *rumen* (ruddy), with 3 additional that are derivatives or compounds: *žućkast* (yellowish), *bledomrk* (pale dark), *bledožut* (pale yellow).

Vehicle - Nouns In extracted similes 356 different nouns appear as vehicles, 209 of them in only one simile. The most frequent nouns and adjectives with which they combine and occur more than once are presented in Table 5. It can be seen that even most frequent nouns combine with just a few adjectives, and that the most frequently used adjectives always exceed $^2/_3$ of all occurrences. Moreover, the noun *krpa* (cloth), in the third row on the list, is used with only one adjective, *bled* (pale).[8] A similar situation is with nouns *led* (ice) and *krv* (blood), which are used with two variants or two synonymous adjectives only, of which one is strongly preferred.

noun	**freq.**	**No.**	**%**	**adjectives**
sneg snow	53	3	90.6	48: beo, 4: čist, 1: nedotaknut 48: white, 4: clean, 1: untouched
smrt death	45	8	82.2	37: bled, 2: lagan, 1: beo, hladan, jak, nem, nepomičan, ukočen 37: pale, 2: light, 1: white, cold, strong, mute, immovable, stiff
krpa cloth	37	1	100.0	37: bled 37: pale
led ice	24	2	83.3	20: hladan, 4: ladan 20: cold, 4: cold
krv blood	18	2	77.8	14: crven, 4: rumen 14: red, 4: ruddy
nebo sky	18	5	66.7	12: plav, 2: vedar, 1: navodnjen, širok, taman 12: blue, 2: clear, 1: watery, wide, dark

Table 5. The most frequent nouns and adjectives with which they combine

[8] It is interesting to note that an analogous simile exists in Greek (Mitrović, Markantonatou, and Krstev 2020), translated to English as “white as cloth”. Hanks (2004) does not mention *cloth* or anything similar among artefacts used in similes in English.

Nouns that refer to animals and plants are often used in similes,[9] although there are no such nouns among the most frequently used nouns in similes presented in Table 5. There are as many as 70 nouns referring to animals, and 34 referring to plants in similes in SrpELTeC, and the most frequent nouns of this kind and adjectives with which they combine are presented in tables 6 and 7.

noun.	**freq.**	**No.**	**%**	**nouns**
jagnje	13	6	46.2	8: miran, 1: blag, dobar, poslušan, smiren, umiljat
lamb				8: calm, 1: mild, good, docile, calm, amiable
ris	12	2	91.7	11: ljut, 1: ljutit
lynx				11: angry, 2: angry
rak	11	1	100	11: crven
crab				11: red
mačka	10	5	40.0	4: oprezan, 3: brz, 1: lagan, pakostan, vešt
cat				4: careful, 3: quick, 1: light, spiteful, dexterous
ovca	10	2	70.0	7: sed, 3: beo
sheep				14: gray-haired, 3: white

Table 6. The most frequent nouns referring to animals and adjectives with which they combine.

One can see that an animal or a plant are sometimes used to denote a specific quality (examples are *ris* (lynx) and *dren* (dogwood)), while sometimes they are associated with various qualities (examples are *mačka* (cat) and *jabuka* (apple)). Animals and plants used in similes are mostly those with which users of Serbian are familiar, although *lav* (lion) and *tigar* (tiger) are also mentioned: *strašan kao lav* (dreadful as a lion) and *brz kao tigar* (quick as a tiger).

As far as proper names are concerned, they appeared in only 4 cases: *dobar kao Hristos* (good as Christ), *crn kao Arapin* (black as an Arab), *brz kao Grk* (quick as a Greek), and *strog kao Turčin* (strict as a Turk).

Marker – conjunctions The conjunction *kao* (as) is by far the mostly used as a marker in similes retrieved from SrpELTeC. Sometimes it is used in a modified form to convey the spoken language in informal speech: *ka'*, *k'o*, *ko*,

[9] Hanks (2004) notes that animals often occur as "secondary subjects" in similes and presents a list of 33 animals appearing in conventionalized similes in English.

noun.	freq.	No.	%	nouns
jabuka	14	5	35.7	5: rumen, 4: pun, 3: zdrav, 1: jedar, okrugao
apple				8: ruddy, 4: plump, 3: healthy, 1: sturdy, round
bor	11	4	36.4	5: prav, 4: visok, 1: zdrav, dičan
pine	11			5: upright, 4: tall, 1: healthy, worthy
dren	7	1	100.0	7: zdrav
dogwood				7: healthy
jela	6	4	33.3	2: prav, vit, 1: vitak, izrastao
fir				2: upright, slender, 1: slender, grown
paprika	5	2	80.0	4: crven, 1: ljut
paprika				4: red, 1: angry

Table 7. The most frequent nouns referring to plants and adjectives with which they combine.

ka, *kano*. Prepositions *nalik na* and *poput* can also be used sometimes, but that was rare in SrpELTeC corpus: we retrieved only two cases using *poput*: *blijed poput krpe* (pale as a cloth) and *raširen poput lepeze* (spread as a fan).

Target The target of the large part of similes extracted from SrpELTeC is a person (see Figure 4, left): a man (335), a woman (184), a child (8), or a group of people (46). A person's appearance (body part) is often referred to, as shown in Figure 4 (right) for all body parts occurring more than once. A person's cloths are mentioned as well: cloths in general (*odelo*, *odeća*, *ruho*) 10 times, *košulja* (shirt) 6 times, *haljina* (dress) 3 times.

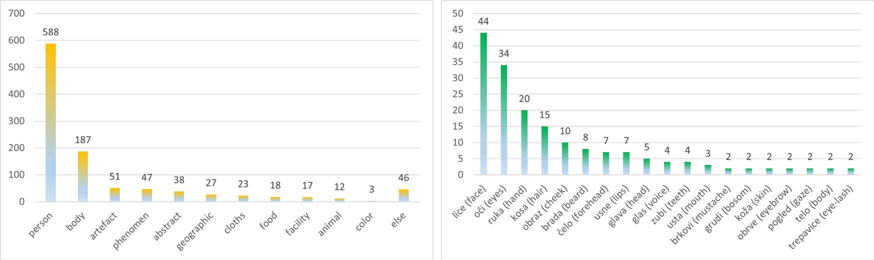


Figure 4. Frequency of targets classified in large groups (left); frequency of body parts used as targets (right).

Extracted data show that some similes are used to describe a large variety of entities. For instance, the most frequent simile *beo kao sneg* (white as snow) connects to many different targets: men, women (individually and collectively), body parts, cloths, animals, household items etc. On the other hand the second on the list, *bled kao krpa* (pale as a cloth) is used only for persons – a man (21), a woman (12), a group of people (2) and a person's face (2), similarly as the third one *bled kao smrt* (pale as death) – a man (20), a woman (14), a group of people (1) and a person's face (2). There are some similes that are used to describe only persons, e.g. *sed kao ovca* (gay-haired as a sheep), while others are used to describe natural or weather phenomena, e.g. *gust kao testo* (thick as dough) is used in SrpELTeC to desribe: *magla* (fog), *mrak*, *pomrčina*, *tama* (darkness), *nebo* (sky), *noć* (night).

Conventionalized similes vs. metaphors One can presume that among similes that occur only once in SrpELTeC there are many which are not conventionalized, or fixed similes, but rather introduced by authors. We tried to analyze the use of this non-conventionalized or infrequent similes by authors who are represented with at least 3 novels in SrpELTEC, and counted the number of these similes among all similes used by a particular author. The results are presented in Figure 5. One can observe that some authors (Svetozar Ćorović and Milutin Uskoković) avoid the use of conventionalized similes, while the others prefer them (Lazar Komarčić and Čedomilj Mijatović). The other authors tend to use both conventionalized and unique similes to a different extent; however, Milan Đ. Milićević, Borisav Stanković and Draga Gavrilović use few similes so it is not possible to establish their preference. On the average, these thirteen authors used almost as many unique similes, as similes shared with other authors – 50.8% of unique similes among all similes used.

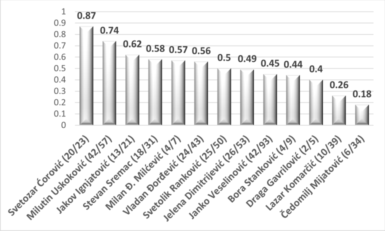


Figure 5. The use of unique similes by authors represented by at least 3 novels in SrpELTeC.

Similes as Verbal MWEs Similes as verbal MWEs consist of, according to (Qadir, Riloff, and Walker 2015), four components: besides the subject of the comparison, the object of the comparison, and the comparator (marker), there is a fourth element, an EVENT, ACT or STATE. Here is an example:

Binko	*ciknu*	*kao*	*guja*	kad ugleda popa.
Binko	squealed	like	a snake	when he saw the priest
TARGET	EVENT	MARKER	VEHICLE	

A GROUND can be optionally included as well; however, it is rare since the intended meaning is usually conveyed without it:

Binko	*ciknu*	(*strašno*)	*kao*	*guja*	kad ugleda popa.
Binko	squealed	(terribly)	like	a snake	when he saw the priest
TARGET	EVENT	(GROUND)	MARKER	VEHICLE	

We have extracted 1,067 verbal similes from SrpELTeC using a regular expression analogous to the one for property similes (method 2 mentioned in Section 2). In this paper we will not analyze this type of similes in depth; we will limit our presentation to only two aspects. The first is the extent to which verbs derived from names of colors are used in verbal similes. The results are presented in Table 8. One can observe that for each color that is not a compound nor a derivative and is used in adjective similes in SrpELTeC, except *siv* (gray), a derived verb participates in verbals similes as well, using in many cases the same vehicles.

Finally, we analyzed to what extent adjective and verbal similes used the same nouns. We found that they have 141 nouns in common. From six most frequently used nouns in adjective similes (Table 5), *led* (ice) and *nebo* (sky) were not used by verbal similes. The other four, *sneg* (snow), *smrt* (death), *krpa* (cloth), *krv* (blood), were used in verbal similes as well, conveying similar meanings, although *sneg* and *smrt* also appear in a different context: *raskraviti se kao sneg* (to loosen up like snow), *kositi kao smrt* (to mow like death), *zvoniti kao smrt* (to ring like death).

All most frequently used animals in adjective similes appear also in verbal similes, conveying the same or similar meaning. For instance, *rak* (crab) is used both in adjective and verbal similes only for its red color. The example of *jagnje* (lamb) is interesting. In adjective similes it is used to characterize someone who is calm, mild, good, etc. While there is no lexical connection between the adjectives and the verbs used, the verbal simile *spavati kao jagnje* (to sleep like a lamb) conveys a similar meaning: only somebody calm, mild, good, etc. can sleep sound.

Plants *jela* (fir) and *dren* (dogwood) do not appear in verbal similes. *Paprika* is used for its red color, *jabuka* (apple) for its red or ruddy color,

adj.	**verbs**	**nouns**
beo	BELETI SE, _pobeleti_	_sneg_*, _kreč_*, _visibaba_
white	to be, to become white	snow, lime, snowdrop
crn	CRNETI SE, _pocrneti_	_zift_*, _zemlja_*, _strnjište_, _Ciganin_
black	to be, to become black	tar, earth, sttuble-field, Gypsy
crven	CRVENETI SE, _pocrvneti_, _zacrveneti se_	_rak_*, _paprika_*, _jabuka_, _ruža_, _trešnja_,...
red	to be, to become red	crab, paprika, apple, rose, cherry,...
plav	PLAVITI SE	_čivit_*
blue	to be blue	indigo
zelen	_pozeleneti_	_trava_*, _žuć_, _gušter_
green	to become green	grass, bile, lizard
žut	ŽUTETI SE, _požuteti_	_limun_*, _vosak_*, _smilje_*, _dukat_, _ćilibar_
yellow	to be, to become yellow	lemon, wax, immortelle, gold coin, amber
modar	MODRETI SE, _pomodreti_	_čivit_*, _more_, _smokva_
livid	to be, to become livid	indigo, sea, fig
rumen	RUMENETI SE, _porumeneti_, _zarumeneti se_	_jabuka_*, _ruža_*, _jagoda_*, _gvožđe_, _žeravica_,...
ruddy	to be, to become pink	apple, rose, strawberry, iron, ember,...

Table 8. Verbs derived from colors used in verbal similes in SrpELTeC; verbs in small caps are imperfective; nouns with an * are also used in adjective similes with the corresponding color.

while *bor* (pine) is used in *porasti kao bor* (to grow up as a pine), meaning to become tall, and consequently upright and slender.

4 Simile in Contemporary Serbian Novels

In order to compare the use of similes in SrpELTeC, which comprises Serbian novels from 1840-1920, with contemporary novels, we compiled an *ad hoc* collection of 22 novels published from 1954 to 2010, which we will call Novels22.[10] The collection contains almost 1.6M words, an average of 72,281 words per novel.[11]

In order to retrieve similes from this collection we used local grammars in the form of finite-state graphs, as explained in Section 2 and in (Krstev, Jaćimović, and Vitas 2020). We enhanced the initial set of local grammars, developed for similes of the incomplete version of SrpELTeC, on the basis of retrieved similes from the complete SrpELTeC, so that now it contains graphs for 557 different similes, while each of these graphs takes care about modifications and variations listed in tables 1 and 2.

All most frequent similes in SrpELTeC (listed in Table 3) except one (*mlad kao kaplja* (young as a drop)) appear also in Novels22, some of them several times: *beo kao sneg* (white as snow), *bled kao krpa* (pale as a cloth), *crven kao krv* (red as blood). However, besides most frequent similes we retrieved in Novels22 some similes that appeared in SrpELTeC only once, e.g. *crn kao Arapin* (black as an Arab) and *lak kao dim* (light as smoke).

The graphs retrieved 69 similes from Novels22, that is, 3.14 per novel (compared to 10.5 in SrpELTeC), or 0.434 per 10,000 words in a novel (compared to 2.20 in SrpELTeC). In interpreting these results one should keep in mind that these graphs can retrieve only similes that were confirmed in SrpELTeC. However, if we consider that, on the average, authors of SrpELTeC used as many unique similes as those used by other authors (Subsection 3.2), an estimate of the average number of similes used by Novels22 authors would be approximately double (6.28), which is still considerably bellow the average use in SrpELTeC. This can suggest that either authors of modern novels

[10] This collection contains 14 novels from the German-Serbian parallel corpus (Andonovski, Šandrih, and Kitanović 2019). The corpus comprises 7 novels originally written in Serbian and 7 novels written in German and translated to Serbian. The remaining 8 novels were taken from the Anthology of Serbian Literature.

[11] According to ELTeC classification the collection comprises 8 short, 12 medium sized, and 2 long novels.

do not use similes as much as their predecessors or that the repertoire of similes has changed over time. In order to confirm or reject either of these hypotheses a systematic research of simile use in contemporary novels is needed.

5 Conclusion

The results presented here will serve as the basis for a future database of similes in Serbian, from which local grammars that recognize and tag them in a text can be automatically produced. Our future work will go in two directions. We will examine the use of simile figures in the Serbian language in literary and other texts by using both general corpora and literary corpora covering different time periods. Also, we will expand our research to similes differing in structure, e.g. *čvrst kao od čelika* (firm as from steel), and to verbal similes like *sevati kao munja* (to blaze as lightening) and *liti kao iz kabla* (to pour as from a bucket).

References

Andonovski, Jelena, Branislava Šandrih, and Olivera Kitanović. 2019. “Bilingual lexical extraction based on word alignment for improving corpus search.” *The Electronic Library.*

Brehmer, Bernhard. 2009. “Äquivalenzbeziehungen zwischen komparativen Phraseologismen im Serbischen und Deutschen.” *Südslavistik online* 1:141–164.

Hanks, Patrick. 2004. “Similes and Sets: the English Preposition like.” In *Jazyky a jazykoveda (Languages and Linguistics : Festschrift for Professor Fr. Čermák,* edited by R. Blatná and V. Petkevič. Prague: Philosophy Faculty of the charles University.

Israel, Michael, Jennifer Riddle Harding, and Vera Tobin. 2004. “On simile.” *Language, culture, and mind* 100.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, Jelena Jaćimović, and Duško Vitas. 2020. "Analysis of Similes in Serbian Literary Texts (1840-1920) Using Computational Methods." In *Proc. of the 4th Int. Conference Computational Linguistics in Bulgaria (CLIB 2020),* edited by Svetla Koeva, 31–41. Sofia, Bulgaria: Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences.

Mitrović, Jelena, Stella Markantonatou, and Cvetana Krstev. 2020. "A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing." In *Multiword Expressions: Drawing on Data from Modern Greek and Other Languages,* edited by Stella Markantonatou and Anastasia Christofidou, 241–262. Research Centre for Scientific Terms / Neologism.

Niculae, Vlad, and Victoria Yaneva. 2013. "Computational considerations of comparisons and similes." In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop,* 89–95.

Qadir, Ashequl, Ellen Riloff, and Marilyn Walker. 2015. "Learning to recognize affective polarity in similes." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 190–200.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "The Serbian Part of the ELTeC Collection – from the Empty List to the 100 Novels Collection." *Infotheca - Journal for Digital Humanities* 21 (2): 7–25. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.1.

Veale, Tony, and Yanfen Hao. 2008. "Enriching WordNet with folk knowledge and stereotypes." In *Proceedings of GWC,* 453–461.

Мршевић-Радовић, Драгана. 1987. *Фразеолошке глаголско-именичке синтагме у савременом српскохрватском језику.* Београд: Филолошки факултет Универзитета у Београду.

SrpELTeC on Platforms: *Udaljeno čitanje*, Aurora, noSketch

UDC 004.738.5:027
DOI 10.18485/infotheca.2021.21.2.7

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs

Petar Popović
petar.popovic@rgf.bg.ac.rs

*University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia*

ABSTRACT: Serbian ELTeC collection (100 novels and extended) developed within COST action CA16204 Distant Reading for European Literary History comprises at this moment 111 novels published in the period 1840-1920. Such a valuable resource is and will be used for various lexical and linguistic research, by using different tools and methodologies. In this paper, three platforms on which these novels are published will be presented: “Udaljeno čitanje”, Aurora and Sketch Engine.

KEYWORDS: distant reading, literary corpus, digital library, concordances, ELTeC.

PAPER SUBMITTED: 2 November 2021
PAPER ACCEPTED: 24 November 2021

1 Introduction

In the past a printed book was the most reliable way to store information and share it with others. Modern digital technology has made it possible to copy, store and share information even from rare, antique and fragile books. The majority of books in the Serbian ELTeC collection (SrpELTeC) were not well known and accessible to the public. Having in mind the effort invested and the importance of the whole collection, we wanted to make it available to as many people as possible. The second aim was to make it available through various channels, in order to meet the needs of different types of users. The second section will present one of the platforms where these novels are published, “Udaljeno čitanje”, intended for readers who would like to see the original print as a picture while reading the digitized version. The Aurora portal, which will be elaborated in the third section, is developed to provide researchers of Serbian literature and other interested users with a detailed insight into the vocabulary of novels, offering them to browse texts, concordances and frequency lists. The Sketch Engine, a platform for corpora management and exploration, as well as for analyzing texts to identify what is typical in a language and what is rare, unusual or emerging usage, which is usually explored by linguists, lexicographers, translators, students and teachers, will be described in the fourth section.

2 udaljenocitanje.unilib.rs

The platform "Udaljeno čitanje"[1] developed at the University library "Svetozar Marković", in cooperation with the University of Belgrade, Faculty of Philology, and the Society for language resources and technologies Jerteh, was supported by national projects[2] in the field of digitization of cultural heritage and contemporary creativity for 2019.

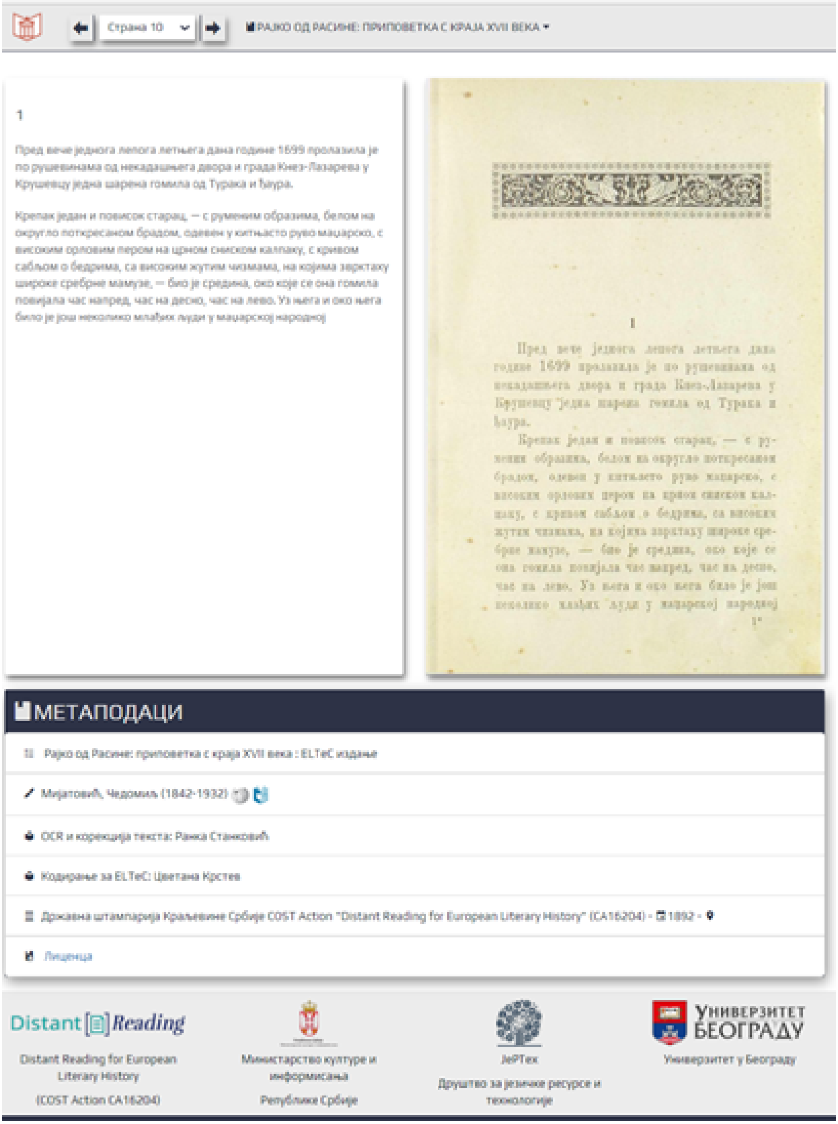
The platform is currently populated with 34 Serbian ELTeC novels, and addition of other Serbian ELTeC novels is planned for the near future. One can browse the novels, select one and read it page by page. A user gets two parallel versions of a chosen text, a picture of the original scanned page on the right and a digitized, machine readable text on the left. Figure 1 presents, in the upper part of the screen, the tenth page of the novel "Rajko od rasine" by Čedomilj Mijatović (SRP18920), with the OCR-ed and corrected text on the left and the original scan of the same page on the right - and at the lower part of the screen, metadata for the same novel. Apart from novel's title, author, publication place, the names of persons responsible for text preparation are given, as well as links to Wikidata, Wikipedia, and Cobiss.[3]

Footnotes in the original text were appropriately encoded and referenced in its digitized version. A small "information" sign in a digitized text signals the existence of a note (a footnote in the original), which, upon a click, appears in the form of a popup window containing the note's text, as can be seen in Figure 2.

[1] Udaljeno čitanje

[2] Project call of the Ministry of Culture and Information of Serbia - Sector for Digitization of Cultural Heritage and Contemporary Creativity for 2019, project number 119-01-00127 / 2019-09 and 401-01-00182 / 2019-09.

[3] Cobiss+



1

Пред вече једнога лепога летњега дана године 1699 пролазила је по рушевинама од некадашњега двора и града Кнез-Лазарева у Крушевцу једна шарена гомила од Турака и ђаура.

Крепак један и повисок старац, — с руменим образима, белом на округло поткресаном брадом, одевен у китњасто руво маџарско, с високим орловим пером на црном свиском калпаку, с кривом сабљом о бедрима, са високим жутим чизмама, на којима зврктаху широке сребрне мамузе, — био је средина, око које се она гомила повијала час напред, час на десно, час на лево. Уз њега и око њега било је још неколико млађих људи у маџарској народној

I

Пред вече једнога лепога летњега дана године 1699 пролазила је по рушевинама од некадашњега двора и града Кнез-Лазарева у Крушевцу једна шарена гомила од Турака и ђаура.

Крепак један и повисок старац, — с руменим образима, белом на округло поткресаном брадом, одевен у китњасто руво маџарско, с високим орловим пером на црном свиском калпаку, с кривом сабљом о бедрима, са високим жутим чизмама, на којима зврктаху широке сребрне мамузе, — био је средина, око које се она гомила повијала час напред, час на десно, час на лево. Уз њега и око њега било је још неколико млађих људи у маџарској народној

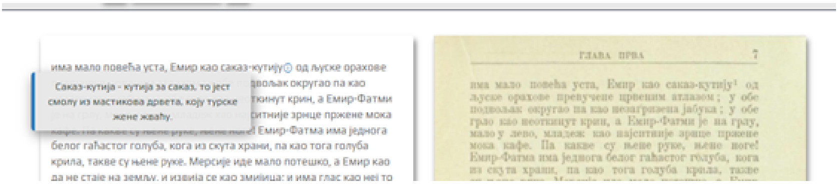
1*

МЕТАПОДАЦИ

- Рајко од Расине: приповетка с краја XVII века : ELTeC издање
- Мијатовић, Чедомиљ (1842-1932)
- OCR и корекција текста: Ранка Станковић
- Кодирање за ELTeC: Цветана Крстев
- Државна штампарија Краљевине Србије COST Action "Distant Reading for European Literary History" (CA16204) - 1892 -
- Лиценца

Distant Reading
Distant Reading for European Literary History
(COST Action CA16204)
Министарство културе и информисања
Републике Србије
ЈеРТех
Друштво за језичке ресурсе и технологије
Универзитет Београду
Универзитет у Београду

Figure 1. The reading layout at udaljenocitanje.unilib.rs



има мало повећа уста, Емир као саказ-кутију① од љуске орахове
подволак округао па као
неоткинут крин, а Емир-Фатми
ситније зрнце пржене мока
Емир-Фатма има једнога
белог гаћастог голуба, кога из скута храни, па као тога голуба
крила, такве су њене руке. Мерсије иде мало потешко, а Емир као
да не стаје на земљу, и извија се као змијица; и има глас као неј то

Саказ-кутија - кутија за саказ, то јест смолу из мастикова дрвета, коју турске жене жваћу.

ГЛАВА ПРВА 7

има мало повећа уста, Емир као саказ-кутију[1] од
љуске орахове превучене црвеним атласом; у обе
подволак округао па као незагрижена јабука; у обе
грло као неоткинут крин, а Емир-Фатми је на грлу,
мало у лево, младеж као најситније зрнце пржене
мока кафе. Па какве су њене руке, њене ноге!
Емир-Фатма има једнога белог гаћастог голуба, кога
из скута храни, па као тога голуба крила, такве

Figure 2. Popup window containing text from an original footnote.

3 Aurora

The name of this portal was chosen to honour the memory of the AURORA[4] (AUtomatska Rutina za Obradu RečnikA – The Automathic Routine for Dictionary Processing) software system for the production of concordances (Vitas 1979), which was the first step in the automatic processing of written texts in the Serbian language. At the home page of this portal, a user can find more information about AURORA program and how it was used to solve problems for which today the corpus processor Unitex/GramLab,[5] a software system that integrates many initial ideas for processing input text in Serbian is being used (Vitas 1980; Krstev 1997, 2008; Vitas and Krstev 2012).

The purpose of the portal is to provide researchers of Serbian literature and other interested users with "microscopic" insight into the vocabulary of a number of works of Serbian literature, both written in prose and verse, offering a user, not only to browse the texts, their concordances and frequency lists, but also to navigate between a text and a list of words extracted from it.

The default preview on the portal's main page shows all titles in two big groups: prose works and poetry. In each of these groups works are listed by authors. The toolbar at the top of the page offers filtering: only names of authors, only female authors and their works, or only works from the ELTeC text collection. Filtering can also be done using the search box, by starting to type an author's name or a work's title, either in Cyrillic or Latin script.

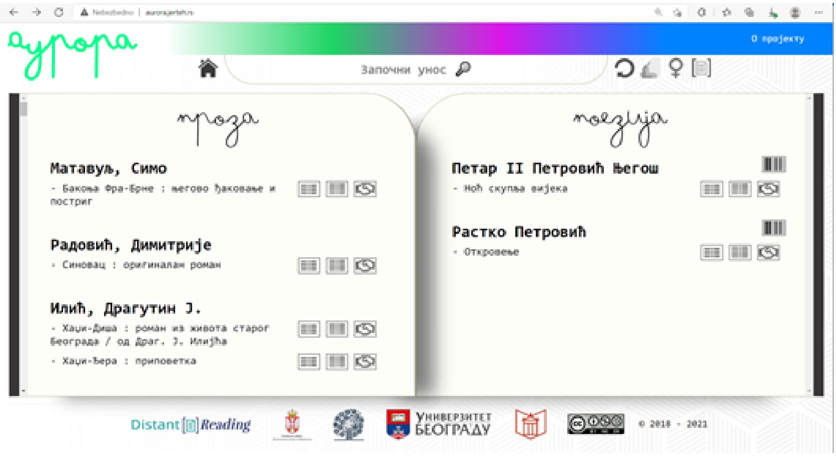
Several authors and titles are linked with Wikidata, while further linking is an ongoing activity and expected to be finished soon. Linking of Wikidata and ELTeC collection is supported by Wikimedia Serbia[6] within the project "wikiELTeC – Wikidata about old Serbian novels from collection ELTeC

[4] Aurora

[5] Unitex Corpus Processing Suite

[6] Wikimedia Serbia

(input, linking of named entities, visualization and analysis)". All 100 novels from Serbian ELTeC sub-collection[7] and 11 from the extended sub-collection are available through the AURORA platform.



Матавуљ, Симо
- Бакоња фра-Брне : његово ђаковање и постриг

Радовић, Димитрије
- Синовац : оригиналан роман

Илић, Драгутин Ј.
- Хаџи-Диша : роман из живота старог Београда / од Драг. Ј. Илијћа
- Хаџи-Ђера : приповетка

Петар II Петровић Његош
- Ноћ скупља вијека

Растко Петровић
- Откровење

Figure 3. The home page of the AURORA platform.

Each text in the AURORA collection is initially processed in order to obtain its inverted version, an alphabetically ordered index in which each entry containing a word and its frequency in a text points to a list of all occurrences of that word. This index can be presented to a user for browsing, ordered either alphabetically or by frequencies. The list can be complete (using button ▤ on the page represented in Figure 3) or filtered, so that the most frequently used words such as conjunctions, prepositions etc. are eliminated (using button ▥). This representation of texts enables the construction of concordances directly and linking of each concordance keyword with a broader context in the full text preview. Namely, using the Unitex-Gramlab locate module with the following regular expression, concordances are generated for all words in a text (except XML elements and their attributes).

```
<WORD><<[^li|div|head|n|p|lg|p\srend=\\\"Tekst\\\"|text|appInfo|
application|encodingDesc|fileDesc|item|encoding|author|body|
```

[7] SrpELTeC

```
document|label|meTypesetSize|publicationStmt|sourceDesc|type|
unknown]>>
```

Another option deletes the so-called stopwords from the list of words:

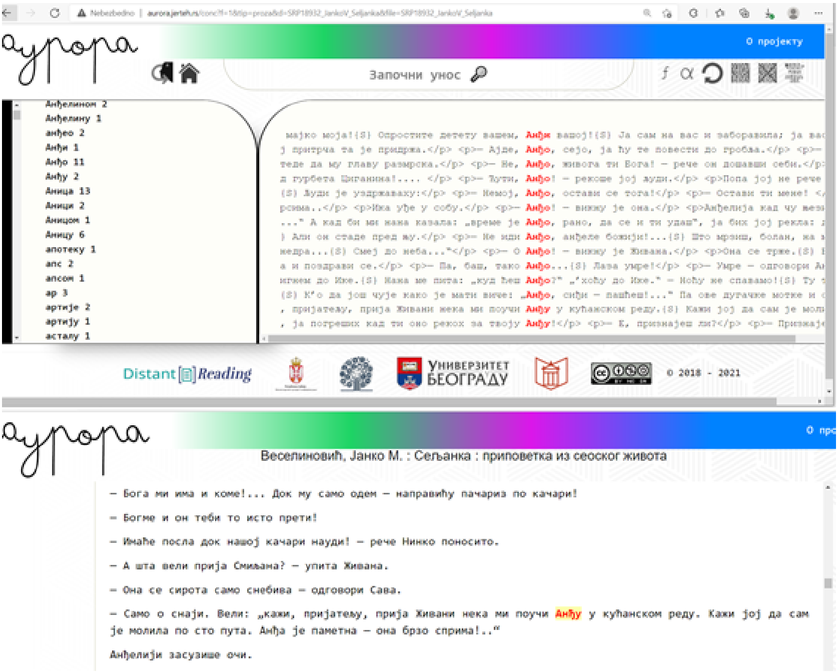
```
<WORD><<[^li|div|head|n|p|lg|p\srend=\\\"Tekst\\\"text|appInfo|
application|encodingDesc|fileDesc|item|encoding|author|body|
document|label|meTypesetSize|publicationStmt|sourceDesc|type|
unknown|и|у|да|на|за|од|а|са|о|из|али|до|што|као|или|по|како|с|када|
јер|због|према|па|после|ако|без|пре|док|око|код|против|него|кад|
уз|већ|где|између|под|пред|преко|међу|иако|кроз|ни]>>"
```

The process of concordance generation is integrated into Leximir (Stanković et al. 2011; Stanković et al. 2012) and for each novel in SrpELTeC that is in level-1 TEI form the following is produced:

- a separate header for metadata extraction;
- a separate body of the text for production of concordances;
- index files (full and reduced);
- html files with concordances (full and reduced);
- html form of the novel.

The full use of the system is illustrated in Figure 4. In this way, AURORA provides insight into the vocabulary of a literary work and is the initial step in creating a dictionary of words used by individual writers. Future versions of this portal, which will use the full content of the system of electronic morphological dictionaries for the Serbian language, will give an even more elaborate insight into literary works.

Let us mention here some directions for future development, one of which will be lemmatizing concordances and associating words in the index with semantic attributes contained in electronic dictionaries. We also plan to integrate named entities, extracted from level-2 version of texts annotated with names of persons and their roles (professions, positions and titles), locations, organisations, events, work titles, and demonyms (Frontini et al. 2020; Šandrih Todorović et al. 2021). This would enable users to browse and search for concordances for a particular named entity class or a particular named entity. Named entities linking with Wikidata and integration with other knowledge bases is also envisaged.



```
Анђелином 2
Анђелину 1
анђео 2
Анђи 1
Анђо 11
Анђу 2
Аница 13
Аници 2
Аницом 1
Аницу 6
апотеку 1
апс 2
апсом 1
ар 3
артије 2
артију 1
асталу 1
```

```
 мајко моја!(S) Опростите детету вашем, Анђи вашој!(S) Ја сам на вас и заборавила; ја вас
ј притрча та је придржа.</p> <p>— Ајде, Анђо, сејо, ја ћу те повести до гробаа.</p> <p>—
теде да му главу размрска.</p> <p>— Не, Анђо, живога ти Бога! — рече он дошавши себи.</p
д гурбета Циганина!.... </p> <p>— Ћути, Анђо! — рекоше јој људи.</p> <p>Попа јој не рече
(S) Људи је уздржаваху:</p> <p>— Немој, Анђо, остави се тога!</p> <p>— Оставих ти мене! </
рсима..</p> <p>Ика уђе у собу.</p> <p>— Анђо! — викну је она.</p> <p>Анђелија кад чу већ
..." А кад би ми мака казала: „време је Анђо, рано, да се и ти удаш", ја бих јој рекла: 
) Али он стаде пред му.</p> <p>— Не иди Анђо, анђеле божији!...(S) Што мрзиш, болан, жа
недра...(S) Смеј до неба..."</p> <p>— О Анђо! — викну је Живана.</p> <p>Она се трже.(S) 
а и поздрави се.</p> <p>— Па, баш, тако Анђо...(S) Лала умре!</p> <p>— Умре — одговори А
иглем до Ике.(S) Мака ме пита: „куд ћеш Анђо?" „'хоћу до Ике." — Нећу ме спавамо!(S) Ту
(S) К'о да још чује како је мати виче: „Анђо, сиђи — пашћеш!..." Па ове дугачке мотке и 
, пријатеву, прија Живани нека ми поучи Анђу у кућанском реду.(S) Кажи јој да сам је моли
, ја погреших кад ти оно рекох за твоју Анђу!</p> <p>— Е, признајеш ли?</p> <p>— Признај
```

— Бога ми има и коме!... Док му само одем — направићу пачариз по качари!

— Богме и он теби то исто прети!

— Имаће посла док нашој качари науди! — рече Нинко поносито.

— А шта вели прија Смиљана? — упита Живана.

— Она се сирота само снебива — одговори Сава.

— Само о снаји. Вели: „кажи, пријатељу, прија Живани нека ми поучи Анђу у кућанском реду. Кажи јој да сам је молила по сто пута. Анђа је паметна — она брзо сприма!..“

Анђелији засузише очи.

Figure 4. The novel *Seljanka* (The Peasant Woman) (SRP18932) by Janko Veselinović: a list of word forms with their frequencies (top left); b) selected concordances for forms of the name *Anđa* (top right); c) one of the chosen forms displayed in full context (bottom).

4 Sketch Engine

Sketch Engine[8] is a widespread tool to explore how language works, based on analysis of corpora compiled from authentic texts of billions of words. It can promptly identify what is typical in language and what is rare, unusual or emerging usage, and enables text analysis and text mining applications through API features.[9] Main end users of Sketch Engine are linguists, lexicographers, translators, students and teachers.

The Sketch Engine contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 60 billion words to provide a truly representative sample of a language. With the Sketch Engine the user can search for a word, phrase or pattern, and results can be presented in the form of word sketches, concordances, word lists, frequency graphs, sketch differences etc (Kilgarriff et al. 2004; Kilgarriff et al. 2014).

A reduced version of the Sketch Engine is available as an open source edition under the name NoSketch Engine. It offers core corpus processing and search features, but it does not support word sketches, preinstalled corpora, term extraction and other more advanced features. A NoSketch Engine node[10] is installed and maintained by the Society for Language Resources and Technologies JeRTeh, offering access to several monolingual and bilingual corpora. For some of them, access is granted to authorized users only, while a number of them are available without authorisation. The SrpELTeC corpus can be freely accessed and searched using CQL (Corpus Query Language).

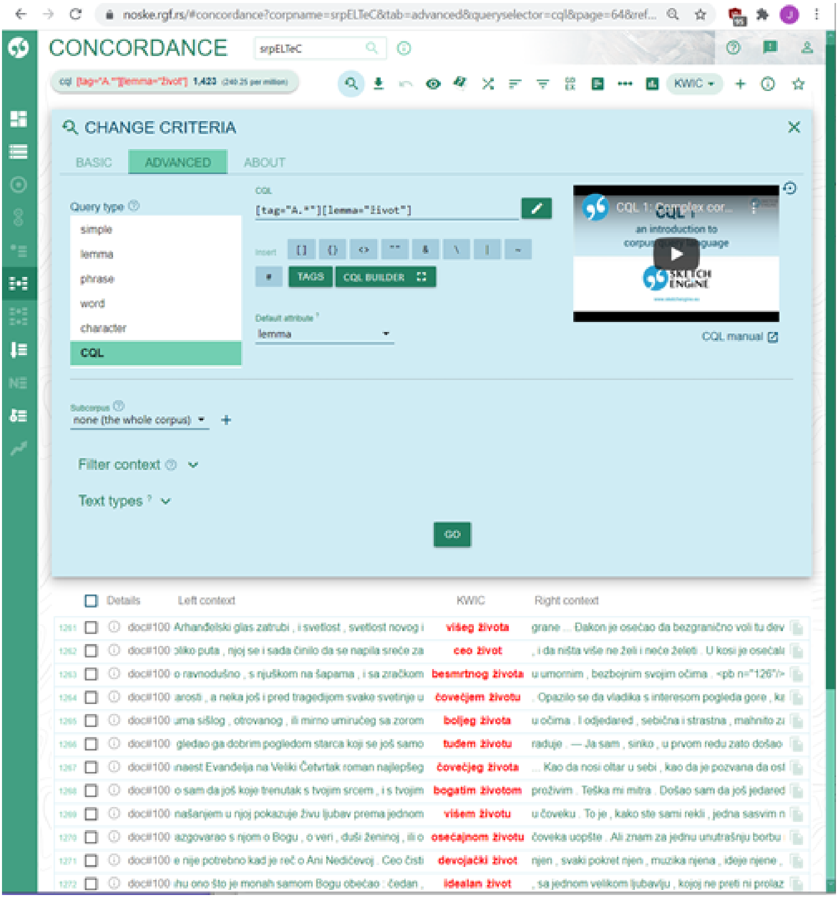
The SrpELTeC corpus in NoSketch is part of speech annotated and lemmatized using TreeTagger (Schmid 1999) with a tagging model, located in the parametric language parameter file, trained on the harmonized resources, which have been manually annotated within different projects (Stanković et al. 2020). The vocabulary that TreeTagger consults when lemmatizing is the system of morphological electronic dictionaries of the Serbian language authored by Cvetana Krstev and Duško Vitas (Krstev 2008; Vitas and Krstev 2012).

Figure 5 presents a page with a simple CQL query `[tag="A.*"][lemma="život"]`, which retrieves concordances with bigrams, where the first word is an adjective and the second is any form of

[8] Sketch Engine

[9] API features

[10] SrpELTeC at JeRTeh



noske.rgf.rs/#concordance?corpname=srpELTeC&tab=advanced&queryselector=cql&page=64&ref...

cql [tag="A.*"][lemma="život"] 1,423 (240.25 per million)

CQL: [tag="A.*"][lemma="život"]

Default attribute: lemma

Subcorpus: none (the whole corpus)

		Details	Left context	KWIC	Right context
1261	□	doc#100	Arhanđelski glas zatrubi , i svetlost , svetlost novog i	višeg života	grane ... Dakon je osećao da bezgranično voli tu dev
1262	□	doc#100	oliko puta , njoj se i sada činilo da se napila sreće za	ceo život	, i da ništa više ne želi i neće želeti . U kosi je osećal
1263	□	doc#100	o ravnodušno , s njuškom na šapama , i sa zračkom	besmrtnog života	u umornim , bezbojnim svojim očima . <pb n="126"/>
1264	□	doc#100	arosti , a neka još i pred tragedijom svake svetinje u	čovečjem životu	. Opazilo se da vladika s interesom pogleda gore , ku
1265	□	doc#100	uma sišlog , otrovanog , ili mirno umirućeg sa zorom	boljeg života	u očima . I odjedared , sebična i strastna , mahnito z
1266	□	doc#100	gledao ga dobrim pogledom starca koji se još samo	tuđem životu	raduje . — Ja sam , sinko , u prvom redu zato došao
1267	□	doc#100	naest Evanđelja na Veliki Četvrtak roman najlepšeg	čovečjeg života	... Kao da nosi oltar u sebi , kao da je pozvana da os
1268	□	doc#100	o sam da još koje trenutak s tvojim srcem , i s tvojim	bogatim životom	proživim . Teška mi mitra . Došao sam da još jedared
1269	□	doc#100	našanjem u njoj pokazuje živu ljubav prema jednom	višem životu	u čoveku . To je , kako ste sami rekli , jedna sasvim n
1270	□	doc#100	azgovarao s njom o Bogu , o veri , duši ženinoj , ili o	osećajnom životu	čoveka uopšte . Ali znam za jednu unutrašnju borbu
1271	□	doc#100	e nije potrebno kad je reč o Ani Nedićevoj . Ceo čisti	devojački život	njen , svaki pokret njen , muzika njena , ideje njene ,
1272	□	doc#100	ihu ono što je monah samom Bogu obećao : čedan ,	idealan život	, sa jednom velikom ljubavlju , kojoj ne preti ni prolaz

Figure 5. Concordances of the query `[tag="A.*"][lemma="život"]` in SrpELTeC.

the lemma *život* (life), e.g. *višeg života* 'higher life', *ceo život* 'whole life', *besmrtnog života* 'immortal life', *boljeg života* 'better life', etc.

The statistics of retrieved KWIC (key words in context) from concordances are presented to the user in the form of a table with absolute and relative (per million) frequencies for lemmatized forms, as presented in Figure 6. The most frequent adjectives that precede the lemma život are: *ceo* 'whole', *nov* 'new', *bračni* 'marital', *društven* 'social', *dug* 'long', etc.

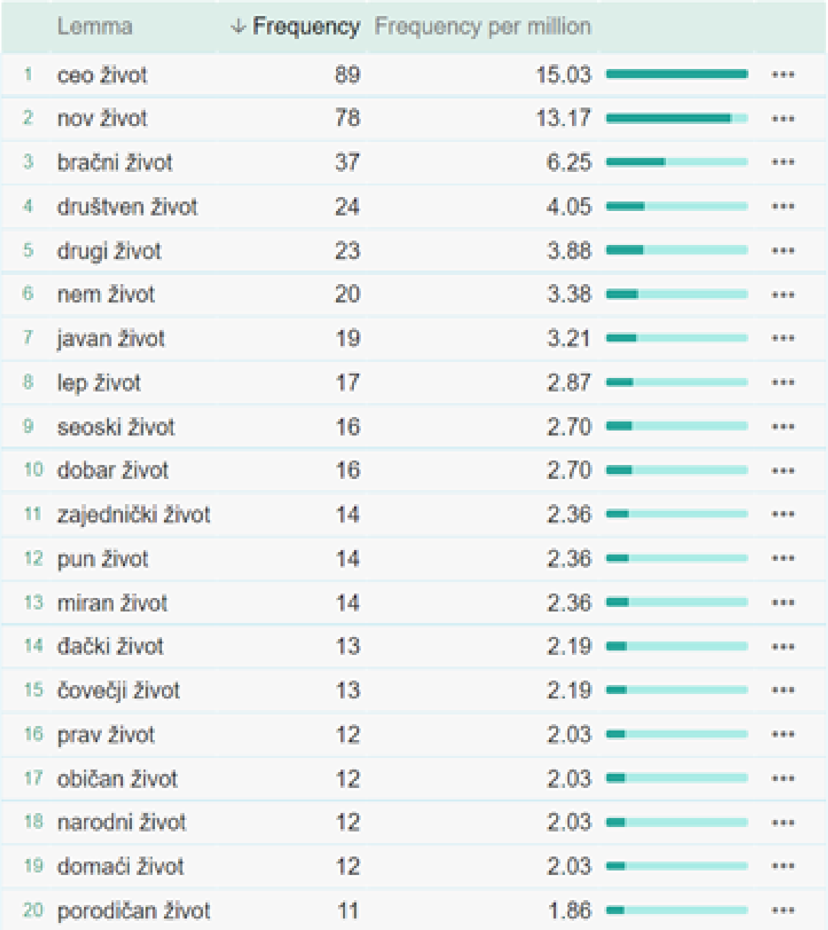
The Faculty of Mining and Geology obtained access to the Sketch engine through the ELEXIS project.[11] Also, the Serbian ELTeC sub-collection is available on this platform for authorized users.[12] As already mentioned, there are additional features available in this environment, such as: word sketches, word clouds, thesaurus, sketch differences etc.

The word sketch feature processes the collocates of a word and other words in its neighborhood (McCarthy et al. 2015; Thomas 2014). Figure 7 presents a word sketch in the form of a set of collocations, grouped by grammatical patterns organized into categories, called grammatical relations, such as words that serve as an object of a verb, words that serve as a subject of a verb, words that modify a word etc. This one-page summary of a word's grammatical and collocational behavior allows for further browsing of concordances for a selected collocate. The sketch grammar is a set of rules written in CQL, based on part of speech tags and regular expressions defining which tokens should be included in the grammatical relation. For example, a subject may be defined as a noun preceding a verb, with additional requirements specified for both components, such as relative positions of the noun and the verb. Also, patterns can include required and optional words between specified components (in this case a noun and a verb).

The visualisation of the word sketch for lemma *život* in SrpELTeC in the form of a diagram is given in Figure 8. Distance from the centre of the big circle in which *život* is located reflects typicality (score): *ceo život* is more typical than *bračan život*. Circle size is related to the frequency: *ceo život* is more frequent than *dug život*. Circle colour indicates which segment (grammatical relation) collocations belong to, because circles may be positioned out of their segments for better visualization. Segment size indicates the size of the grammatical relation relative to other visualized relations, i.e. the number of collocations it contains in total, not just the number of collocations that are visualized.

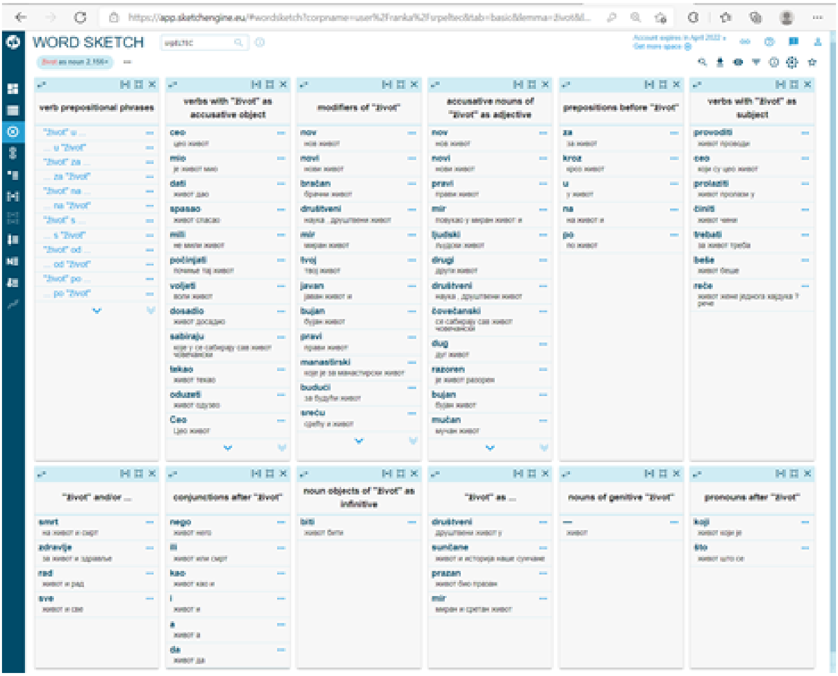
[11] European lexicographic infrastructure

[12] SrpELTeC at Sketch Engine



	Lemma	↓ Frequency	Frequency per million
1	ceo život	89	15.03
2	nov život	78	13.17
3	bračni život	37	6.25
4	društven život	24	4.05
5	drugi život	23	3.88
6	nem život	20	3.38
7	javan život	19	3.21
8	lep život	17	2.87
9	seoski život	16	2.70
10	dobar život	16	2.70
11	zajednički život	14	2.36
12	pun život	14	2.36
13	miran život	14	2.36
14	đački život	13	2.19
15	čovečji život	13	2.19
16	prav život	12	2.03
17	običan život	12	2.03
18	narodni život	12	2.03
19	domaći život	12	2.03
20	porodičan život	11	1.86

Figure 6. The frequency results of the query `[tag="A.*"][lemma="život"]` on SrpELTeC.



https://app.sketchengine.eu/#wordsketch?corpname=user%2Franka%2Fsrpeltec&tab=basic&lemma=život&l...

WORD SKETCH srpELTeC

život as noun 2,156×

Account expires in April 2022 • Get more space

verb prepositional phrases
- "život" u ...
- ... u "život"
- "život" za ...
- ... za "život"
- "život" na ...
- ... na "život"
- "život" s ...
- ... s "život"
- "život" od ...
- ... od "život"
- "život" po ...
- ... po "život"

verbs with "život" as accusative object
- ceo — цео живот
- mio — је живот мио
- dati — живот дао
- spasao — живот спасао
- mili — не миле живот
- počinjati — почиње тај живот
- voljeti — воли живот
- dosadio — живот досадио
- sabiraju — који у се сабирају сав живот човечански
- tekao — живот текао
- oduzeti — живот одузео
- Ceo — Цео живот

modifiers of "život"
- nov — нов живот
- novi — нови живот
- bračan — брачни живот
- društveni — наука , друштвени живот
- mir — миран живот
- tvoj — твој живот
- javan — јаван живот и
- bujan — бујан живот
- pravi — прави живот
- manastirski — који је за манастирски живот
- budući — за будући живот
- sreću — срећу и живот

accusative nouns of "život" as adjective
- nov — нов живот
- novi — нови живот
- pravi — прави живот
- mir — повукао у миран живот и
- ljudski — људски живот
- drugi — други живот
- društveni — наука , друштвени живот
- čovečanski — се сабирају сав живот човечански
- dug — дуг живот
- razoren — је живот разорен
- bujan — бујан живот
- mučan — мучан живот

prepositions before "život"
- za — за живот
- kroz — кроз живот
- u — у живот
- na — на живот и
- po — по живот

verbs with "život" as subject
- provoditi — живот проводи
- ceo — који су цео живот
- prolaziti — живот пролази у
- činiti — живот чини
- trebati — за живот треба
- beše — живот беше
- reče — живот мене једнога кадгод ? рече

"život" and/or ...
- smrt — на живот и смрт
- zdravlje — за живот и здравље
- rad — живот и рад
- sve — живот и све

conjunctions after "život"
- nego — живот него
- ili — живот или смрт
- kao — живот као и
- i — живот и
- a — живот а
- da — живот да

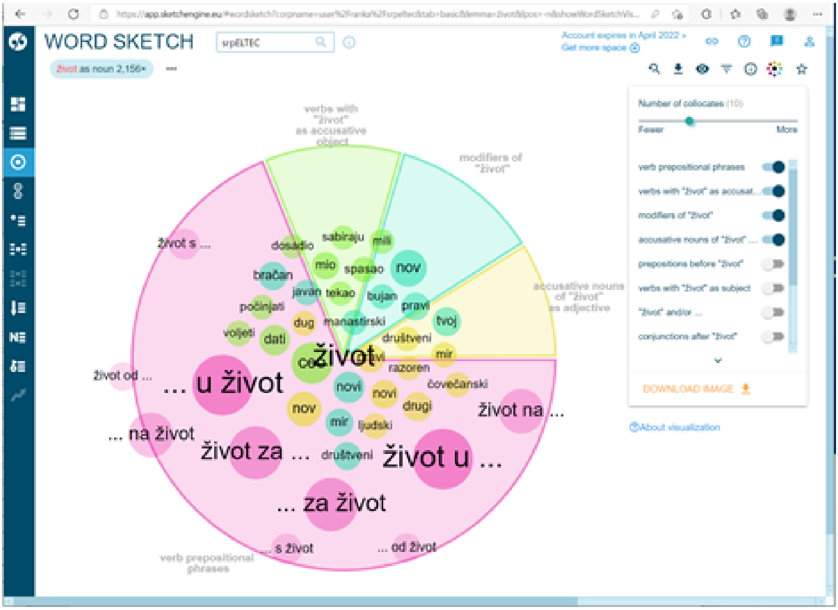
noun objects of "život" as infinitive
- biti — живот бити

"život" as ...
- društveni — друштвени живот у
- sunčane — живот и историја наше сунчане
- prazan — живот био празан
- mir — миран и срећан живот

nouns of genitive "život"
- — — живот

pronouns after "život"
- koji — живот који је
- što — живот што се

Figure 7. One-page overview of the word sketch for the lemma *život* in srpELTeC.



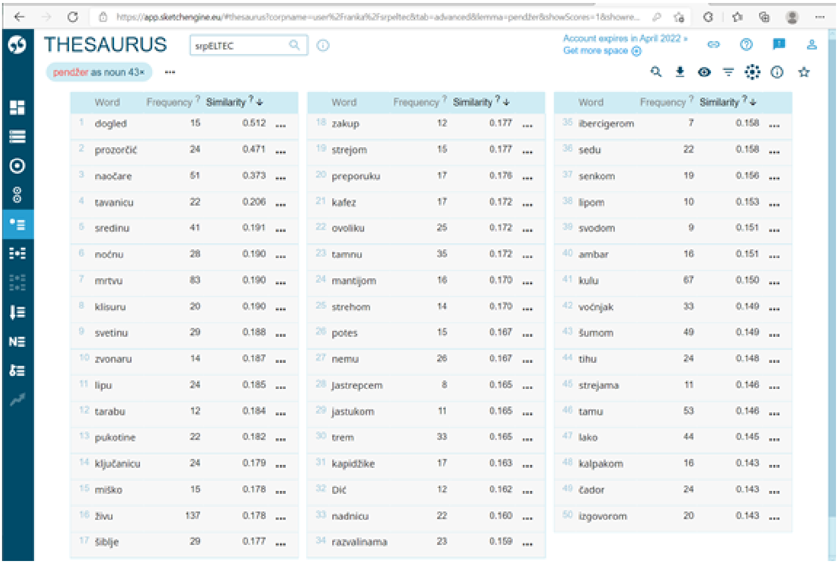
Number of collocates (10)
Fewer — More
verb prepositional phrases ●
verbs with "život" as accusat... ●
modifiers of "život" ●
accusative nouns of "život" ... ●
prepositions before "život" ○
verbs with "život" as subject ○
"život" and/or ... ○
conjunctions after "život" ○

Figure 8. The visualization of word sketch for lemma *život* in SrpELTeC.

The thesaurus presented in Figure 9 is an automatically generated list of synonyms or words belonging to the same semantic field as the required word *život*. The thesaurus entries are retrieved from the context in which the word *život* appears in the srpELTeC corpus.

The synonym candidates are identified automatically from the context in which they occur, relying on the distributional semantic theory hypothesis according to which words that appear in the same context have similar meaning. In the Sketch Engine interpretation, it means that words with similar collocations probably have similar meaning, so the word sketch serves as the base for the calculation of the similarity score. To determine synonyms for *život*, the word sketches of all nouns are compared with its word sketch, and those that share the largest proportion of collocates are listed as similar words. The score assigned to the synonym represents the percentage of shared collocates.

Visualisation in Figure 10 contains information about the frequency and similarity scores of the lemma *život* in SrpELTeC. A circle size reflects the frequency of the encircled word: *čovek* 'man' is more frequent than *rad* 'work'.


https://app.sketchengine.eu/#thesaurus?corpname=user%2Franka%2Fsrpeltec&tab=advanced&lemma=pendžer&showScores=1&showre...
THESAURUS
srpELTEC
pendžer as noun 43×
Account expires in April 2022 »
Get more space

	Word	Frequency	Similarity
1	dogled	15	0.512
2	prozorčić	24	0.471
3	naočare	51	0.373
4	tavanicu	22	0.206
5	sredinu	41	0.191
6	noćnu	28	0.190
7	mrtvu	83	0.190
8	klisuru	20	0.190
9	svetinu	29	0.188
10	zvonaru	14	0.187
11	lipu	24	0.185
12	tarabu	12	0.184
13	pukotine	22	0.182
14	ključanicu	24	0.179
15	miško	15	0.178
16	živu	137	0.178
17	šiblje	29	0.177
18	zakup	12	0.177
19	strejom	15	0.177
20	preporuku	17	0.176
21	kafez	17	0.172
22	ovoliku	25	0.172
23	tamnu	35	0.172
24	mantijom	16	0.170
25	strehom	14	0.170
26	potes	15	0.167
27	nemu	26	0.167
28	jastrepcem	8	0.165
29	jastukom	11	0.165
30	trem	33	0.165
31	kapidžike	17	0.163
32	Dić	12	0.162
33	nadnicu	22	0.160
34	razvalinama	23	0.159
35	ibercigerom	7	0.158
36	sedu	22	0.158
37	senkom	19	0.156
38	lipom	10	0.153
39	svodom	9	0.151
40	ambar	16	0.151
41	kulu	67	0.150
42	voćnjak	33	0.149
43	šumom	49	0.149
44	tihu	24	0.148
45	strejama	11	0.146
46	tamu	53	0.146
47	lako	44	0.145
48	kalpakom	16	0.143
49	čador	24	0.143
50	izgovorom	20	0.143

Figure 9. Thesaurus for lemma *život* in SrpELTeC.

The distance from the circle centre, where again *život* is situated, depends on the similarity score: *dan* 'day' is more similar to *život* than *posao* 'work'. This example shows that none of the candidates need actually be synonyms. To obtain more reliable results one would need to work on a much bigger corpus.

The word sketch difference provides comparisons by contrasting collocations that can be retrieved by lemmas, word forms or subcorpora. Comparing the collocations can provide a deeper understanding of the difference in use and meaning. Figure 11 presents the word sketch difference for *život* and *smrt* (death), which compares the use of the two lemmas by comparing their collocates. Different colours are assigned to the chosen words and their word sketches, and for each collocate, in each grammatical relation separately, the results for both words are compared. The colours indicate the word for which the collocate is more frequent (blue for *život*, pink for *smrt*), while the shade of the colour indicates the strength of the collocation. The words in white (for example *ja*, *sam* in the centre of the bottom boxes in Figure 11) are collocates without a preference.

An additional option is the "word forms" option, which compares the use of two different word forms of the same lemma via their collocates. A third

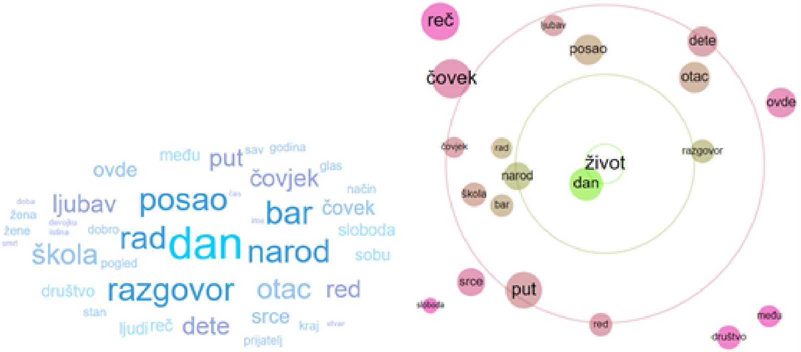
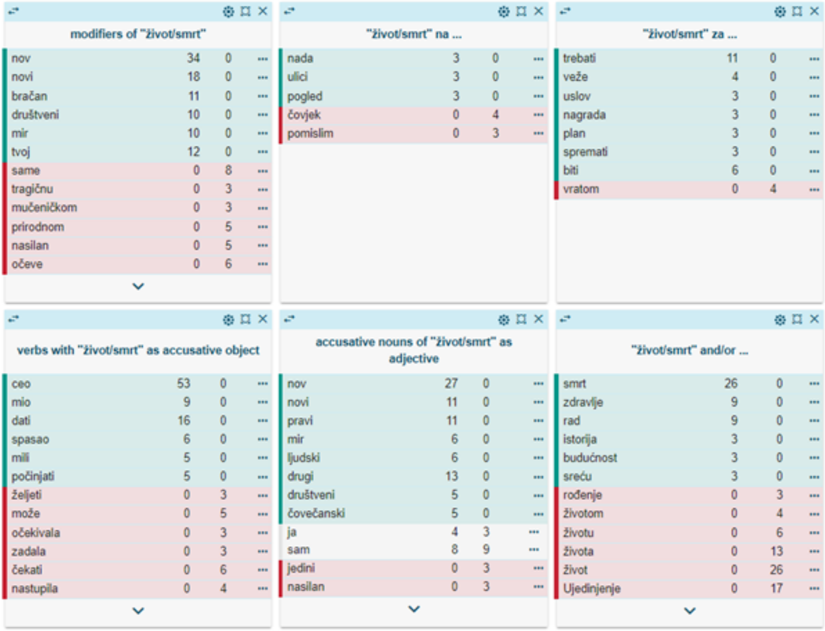


Figure 10. Visualisation of the thesaurus for lemma *život* in SrpELTeC.

option is "subcorpora", which compares the use of the same lemma in two different subcorpora of the same corpus, via their collocates.

5 Concluding remarks

In this paper three platforms supporting the SrpELTeC collection of novels are presented. The first is "Udaljeno čitanje", which enables browsing and reading of digitized text, with a preview of the scanned original. The platform will be further improved by publishing more novels, as well as by introducing advanced features for search and filtering. The second platform is the Aurora portal, which provides researchers with "microscopic" insight into the vocabulary of selected Serbian literary works. In addition to the text, concordances and word frequencies are also available, as well as navigation between a text and a list of words extracted from it. Apart from further expansion of resources, Aurora will be more tightly integrated with Wikidata, presenting results from predefined SPARQL queries, with authors and their novels, novels linked with main characters and their roles, in form of graphs, tables, timelines (Ikonić Nešić, Stanković, and Rujević 2021) and locations of the events on the maps etc. Browsing lists of rare words and browsing by authors, will also be enabled. The third platform is the Sketch Engine, for corpora management and exploration, as well as for analyzing large texts. Integration of the Aurora portal and the Sketch Engine is envisaged, since Aurora is not optimised for large novels. We believe that the developed platforms will contribute to raising the visibility of SrpELTeC, a valuable resource in Serbian language for linguists, but also bring a part of literary history that was unknown or unavailable closer to to the wider community.



modifiers of "život/smrt"		
nov	34	0
novi	18	0
bračan	11	0
društveni	10	0
mir	10	0
tvoj	12	0
same	0	8
tragičnu	0	3
mučeničkom	0	3
prirodnom	0	5
nasilan	0	5
očeve	0	6

"život/smrt" na ...		
nada	3	0
ulici	3	0
pogled	3	0
čovjek	0	4
pomislim	0	3

"život/smrt" za ...		
trebati	11	0
veže	4	0
uslov	3	0
nagrada	3	0
plan	3	0
spremati	3	0
biti	6	0
vratom	0	4

verbs with "život/smrt" as accusative object		
ceo	53	0
mio	9	0
dati	16	0
spasao	6	0
mili	5	0
počinjati	5	0
željeti	0	3
može	0	5
očekivala	0	3
zadala	0	3
čekati	0	6
nastupila	0	4

accusative nouns of "život/smrt" as adjective		
nov	27	0
novi	11	0
pravi	11	0
mir	6	0
ljudski	6	0
drugi	13	0
društveni	5	0
čovečanski	5	0
ja	4	3
sam	8	9
jedini	0	3
nasilan	0	3

"život/smrt" and/or ...		
smrt	26	0
zdravlje	9	0
rad	9	0
istorija	3	0
budućnost	3	0
sreću	3	0
rođenje	0	3
životom	0	4
životu	0	6
života	0	13
život	0	26
Ujedinjenje	0	17

Figure 11. The word sketch difference for *život* (blue if preferred) and *smrt* (pink if preferred) in SrpELTeC.

Acknowledgment

The platform Udaljeno čitanje is developed with the support of projects in the field of digitization of cultural heritage and contemporary creativity for 2019: 119-01-00127 / 2019-09 and 401-01-00182 / 2019-09, supported by the Ministry of Culture and Information of Serbia – Sector for Digitization of Cultural Heritage and Contemporary Creativity.

Linking of Wikidata and ELTeC collection is supported by Wikimedia Serbia within the project "wikiELTeC – Wikidata about old Serbian novels from collection ELTeC (input, linking of named entities, visualization and analysis)".

The text collection preparation is supported by the COST Action 16204 – Distant Reading for European Literary History support. Access to SketchEngine is provided by the ELEXIS project funded by the European Union's Horizon 2020 research and innovation program under grant number 731015.

References

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020.*

Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. "Serbian ELTeC Sub-collection in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (2): 60–87. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.4.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "Itri-04-08 the sketch engine." *Information Technology* 105 (116).

Krstev, Cvetana. 1997. "Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije." PhD diss., Univerzitet u Beogradu, Matematički fakultet, September.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology of the University of Belgrade.

McCarthy, Diana, Adam Kilgarriff, Milos Jakubicek, and Siva Reddy. 2015. "Semantic word sketches." *Corpus Linguistics 2015.*

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER&Beyond: The Archaic and the Modern Intertwined." In *Deep Learning Natural Language Processing Methods and Applications – Proc. of the Int. Conf. Recent Advances in Natural Language Processing (RANLP 2021),* edited by Galia Angelova et al., 1252–1260. INCOMA Ltd. https://doi.org/10.26615/978-954-452-072-4_141.

Schmid, Helmut. 1999. "Improvements in part-of-speech tagging with an application to German." In *Natural language processing using very large corpora,* 13–25. Springer.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. "A tool for enhanced search of multilingual digital libraries of e-journals." In *Proc. of the 8th LREC),* edited by Nicoletta Calzolari et al. Istanbul, Turkey: European Language Resources Association (ELRA).

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference, October 2011, Jachranka, Poland,* 77–84.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proc. of The 12th LREC,* 3947–3955. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.487.

Thomas, James. 2014. "Discovering English with the Sketch Engine." *Research-publishing. net,* 161–176.

Vitas, Duško. 1979. "Prikaz jednog sistema za automatsku obradu teksta." In *INFORMATICA'79, Bled,* 7101 1–5.

Vitas, Duško. 1980. "Generisanje imeničkih oblika u srpskohrvatskom." *Informatica,* no. 3, 49–55.

Vitas, Duško, and Cvetana Krstev. 2012. "Processing of Corpora of Serbian Using Electronic Dictionaries." *Prace Filologiczne* XVIII:279–292.

OCR and TEI for the production of ELTeC – Würzburg Training School, 16-17 April 2018

Jelena Andonovski
andonovski@unilib.rs
University Library
"Svetozar Marković"
Belgrade, Serbia

PAPER SUBMITTED: 23 June 2021
PAPER ACCEPTED: 07 November 2021

As the basic task of the Action CA16204 D-Reading is corpus preparation, from the beginning of the project the most important issue was to define methods for text processing, annotation of corpus material and metadata creation. With these aims in mind, the project coordinators organized the first workshop for the Action partners. The two-day workshop was held on April 16 and 17, 2018, at the University of Würzburg, Germany (Figure 2).[1] The organizers were Leonard Konle and Fotis Jannidis from the University of Würzburg, while lecturers were Leonard Konle and Christian Reul, also from the University of Würzburg, and Lou Burnard, an internationally recognised expert in digital humanities, particularly in the area of text encoding and digital libraries. The workshop was attended by 11 participants from 10 countries: Jelena Andonovski from the University of Belgrade, Serbia; Alex Ciorogar from UBB Cluj-Napoca, România; Simon Gabay from the University of Neuchatel, Switzerland; Meliha Hadžić from the Burch University, Sarajevo, Bosnia and Herzegovina; Magdalena Krol from the Institute of Polish Language, Polish Academy of Sciences, Poland; Ioana Lionte from the University "Alexandru Ioan Cuza", Iași, Romania; Anna Rehorkova from the Charles University, Prague, Czech Republic; Floriana Sciumbata from the University of Trieste, Italia; Anna Maria Sichani from the Kings Digital Lab, London, Great Britain; Adeliana Silva from the Nova University of Lisbon, Portugal; Andrejka Zejn from the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia (Figure 1).

The aim of the workshop was to introduce participants to methods of text processing and annotation.[2] Methods for Optical Character Recognition (OCR) were presented, as well as the available software for that purpose, and then it was explained how to encode texts in the electronic format chosen

[1] About the training school

[2] Training material

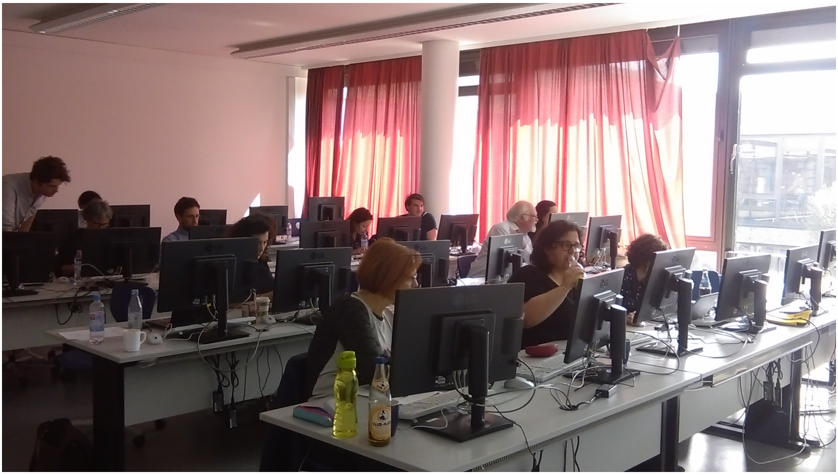


Figure 1. The second day of the workshop

for ELTeC,[3] that is TEI/XML. The first day of the workshop, April 16, was dedicated to OCR. First of all, lecturers presented the basic characteristics of OCR in general, and then some software packages, Abbyy FineReader[4] and an open-source tool OCR4all.[5] Thomas M. Breuel from the University of Kaiserslautern/DFKI, Xerox, Google, currently Nvidia, presented OCR4all and pointed out some of its basic characteristics:

1. It was primarily created to digitally explore very early printed texts;
2. It was prepared to be an open source tool;
3. It was made to be understandable and adaptable for users lacking technical experience;
4. It is independent from the software platform;
5. It is based on some open-source tools (the central part is the OCRopus tool based on Python, which enables preprocessing, layout segmentation, character recognition and model training).

The first day ended with the lecture given by Anna Řehořková from the Institute of the Czech National Corpus, who shared the Institute's experiences in digitizing material for the Czech National Corpus. During the day,

[3] About ELTeC at DH2019

[4] Abbyy FineReader

[5] OCR4all

Figure 2. Participants of the workshop enjoying Würzburg

there was one lunch break and one coffee break, and in the evening, a dinner was organized for all participants, during which they could exchange personal experiences in digitization and corpus preparation.

The second workshop day, April 17, was dedicated to corpus annotation and metadata creation. The lecturer was Lou Burnard. At the beginning, he presented the XML Editor oXygen and its characteristics, then TEI/XML structure for the text encoding and at the end TEI header[6] for metadata creation. After that, he introduced participants to the specially prepared ELTeC encoding Schemas. He explained the method of ELTeC encoding Schemas' creation according to TEI P5 Guidelines, the ODD chaining technique - One Document Does it all.[7] In this way three levels of text encoding were prepared:

1. **Level 0** (eltec-0) – basic TEI structure for text encoding in ELTeC corpus;
2. **Level 1** (eltec-1) – additional elements for encoding (for example, text annotation for lyrics);
3. **Level 2** (eltec-2) – linguistic and semantic annotation of texts, at the level of individual tokens and segments.

[6] TEI header

[7] ODD chaining technique - One Document Does it all

After the lunch break participants had a practical work session, during which they worked with a concrete example: they encoded previously OCRed text using TEI/XML and checked its correctness with the XML Editor oXygen using XML Schemas eltec-0 and eltec-1.

In order to prepare metadata, participants were introduced to the TEI Header. At that moment, the structure of metadata for ELTeC corpus was not yet precisely defined. The workshop was thus an excellent place to discuss the header structure, having in mind that participants were librarians, literary scholars, as well as researchers working in digital humanities. Many different opinions could be heard about metadata creation, which data are important, which have to be mandatory etc. At the end of the workshop an assignment was prepared for participants, to complete when they return home: to prepare one text for the ELTeC corpus according to the guidelines they were given at the workshop.

Workshop "Methods and Tools of Distant Reading Adapted to Multiple European Languages" at the Galway Training School

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia

PAPER SUBMITTED: 11 October 2021
PAPER ACCEPTED: 10 November 2021

Galway training school[1] was organized within the COST[2] action CA16204, project *Distant Reading for European Literary History*.[3] One of the main goals of this action was to create a network of researchers which would develop resources and methods within the Distant Reading paradigm – the usage of computational methods in the analysis of large amounts of literary texts. In order to achieve this, one of the action's outcomes is the preparation of a large literary corpus in several European languages, while another is the establishment of good, innovative methods and techniques for computational analysis of these specific texts.

In accordance with that, a training school was organized, and within it two workshops: a workshop on methods and tools of distant reading and a workshop on theoretical concepts and their confrontation with computational methods, with a goal of familiarizing participants with this topic. This training school, the second one organized by this action, was held at the premises of the National University of Ireland in Galway in December, 2018. The aim of the first workshop on methods and tools of distant reading was to present to participants a set of tools that can be applied for solving certain tasks within the distant reading paradigm.

The first tool presented was TXM,[4] which is used for corpus management and analysis using methods of textometry. The tool was presented by Serge Heiden from the *École normale supérieure de Lyon*, one of the authors of the software. From the corpus management point of view, TXM can create

[1] Galway Training School 2018 General Information

[2] COST Actions

[3] D-reading

[4] textometrie

corpora from text files, supporting both plain and annotated texts in different formats, such as XML, including TEI. Corpus-related metadata can be added from a separate XML file during the text import. Once created, the corpus can be browsed, searched and analyzed using various statistical and textometry methods, where values such as absolute and relative frequencies for word, word type and lemma can be calculated and exported for simple or complex queries. Software can also perform search for factorial correspondences, hierarchical classification, collocation analysis, etc. Participants had the practical task to create a corpus and then analyze the mentioned metrics using the TXM software.

The second theme of the workshop was topic modeling, and how it can be done using the *TopicsExplorer* package.[5] The lecturer was Steffen Pielström from the Würzburg University. Participants were introduced with topic modeling – grouping texts or documents by topic and were given the task to try it out using some texts and *TopicsExplorer* software. Participants were to experiment, changing the hyperparameters (number of topics searched, lists of stop words, etc.) and analyze the different results obtained in this way. After the experiments, participants discussed what the analyzed texts were about, what their key words were, and which texts are thematically similar and why.

The lecturer of the third course related to network analysis was Meliha Handžić from the *International University BURCH* in Sarajevo. Within this course, two software packages for visualizing text networks were presented. *Palladio*,[6] a software developed by a team from Stanford University that provides spacetime labeling and visualization of textual sources on an interactive map, and *Gephi*,[7] which is used for two-dimensional visualization of distances between the entities, in this case texts or authors. Distance values can represent any parameter and are imported in the form of a matrix table showing the mutual distance of all nodes. In addition to the predefined output, *Gephi* also offers the application of various graph transformation algorithms, to procure truer or better-looking images.

The last course, held on the last day of the training school, was devoted to the *stylo* package[8] for the programming language R, and the lecturer was Joanna Byszuk from the Polish Academy of Sciences – Polish language institute in Krakow, one of the institutions that developed the package. Par-

[5] TopicsExplorer
[6] Palladio
[7] Gephi
[8] stylo

ticipants were introduced with some possibilities that the package provides, such as stylometric analysis of texts and authors, finding of stylometric similarities and differences between documents, document clustering and dealing with the problem of authorship attribution.[9] Participants were required to create a set of documents for analysis, run *stylo* on their computer and analyze the texts. The corpus of short English novels was provided for participants that did not have access to a corpus of texts in their mother tongue. The last task for participants in this course was to experiment, using the imported corpus, with training and testing a computer model in order to determine document authorship.

Upon completion of this training school, participants were able to prepare a corpus, analyze it, present results, and were thus ready to do research and to study literary works through the distant reading paradigm.

[9] Stylo in Galway

Budapest Training School – Canonization in Distant Reading Research

Vasilije Milnović
milnovic@unilib.rs
University Library
"Svetozar Marković"
Belgrade, Serbia

PAPER SUBMITTED: 10 November 2021
PAPER ACCEPTED: 13 November 2021

As a part of the COST action "Distant Reading for European Literary History" (COST Action CA16204), the third Training School was organized at the Centre for Digital Humanities – Eötvös Loránd University in Budapest from September 23 to 25. Within this Training School, participants could choose to attend one of three parallel tracks: Corpus design and text contribution for ELTeC, Natural Language Processing for Distant Reading and Canonization in Distant Reading Research.

Since my scientific expertise is related to the study of the relationship between tradition and avant-garde and a certain re-examination of canons in the traditional sense, I chose to attend the last track. I did not have real experience in using track-related digital tools. That was the reason why this kind of training was of great use to me to get acquainted with methodologies I might use for my future scientific work.

Lecturers within the chosen topic were:

- Marijan Dović (Slovenia, Literature History), who gave an introductory lecture on the topic "Literary Canon: From Traditional to Contemporary Approaches",
- Maciej Eder (Poland, Linguistics), who gave guidelines for the application of Digital Humanities and Distance Reading in literary language research,
- Karina van Dalen-Oskam (Netherlands, Digital Humanities), who gave a presentation "From the Riddle of Literary Quality to the Riddle of the Literary Canon: a meta-perspective" and
- Christof Schoch (Germany, Digital Humanities), the Coordinator of the entire COST Action, with the topic of technical details of Distant Reading research "Operationalization, Formalization, Modeling: Measuring Degrees of Canonization for Distant Reading Research".

After the importance of a kind of meta-perspective of Distant Reading and the possibility of applying such a tool to research the canonicity of literary works was pointed out, participants were introduced to the many possibilities of digital research of literary text. Regardless of whether the research is conducted according to keywords, or digital tools are used in terms of some type of synthesis, digital humanities provide various possibilities that can also serve as a form of re-examination of existing canons. Distant Reading can be used for a new approach to interdisciplinary analysis of literary texts from the past, including those from the Serbian language collection, for the simple reason that the software can detect various aspects of language in specific texts, which cannot be observed by simple close reading.

In this sense, Distant Reading uses technology to get a bird's eye view of a corpus, and represents one of the most important resources for research in the field of linguistics and related language disciplines. It is used in all linguistic disciplines as a tool for literary and artistic texts. In such research, a special place is occupied by the statistical analysis of language, which can provide conclusions that cannot be drawn on the basis of close reading. As an illustration of distant reading methods, some of the opportunities regarding the use of certain words in the novels of Charles Dickens and Jane Austen, and their connection with existing language corpora, were presented to the participants of the training school.

Within all lectures, participants were presented with basic information about the traditional canon (literary and cultural), as well as the application of the philosophy of distant reading to study the canonicity of literary works. While reaching conclusions about both distant and close reading, an important methodological question for the digital humanities was raised during the discussion: what function might (canonical) close reading fulfil in the digital analysis of large corpora? Is there a way of combining close and distant reading – which many scholars have argued for — without seemingly undermining the necessity of one or the other?

To answer this question, participating in the COST action was very beneficial, because the Serbian ELTeC sub-collection now contains 100 novels, from the period 1840-1920, with more in its extension, including many works that were not part of the official national cultural canon. Both types of research can now be done on this sample - close or distant reading - and based on everything I have learned in this workshop – I am sure that it is possible to apply them both in studying the canonicity of a work. This possibility is especially well expressed when it comes to the avant-garde, the main topic of my scientific research, because even in the so-called developed or large cultures it is part of an alternative or underground canon. Therefore, I am very interested in spreading my knowledge on the topic of distant reading and its possible application in scientific research.

Short Term Scientific Mission to Krakow: Comparative Stylistic and Morphosyntactic Analysis of ELTeC Texts Using Stylo R Package

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia

PAPER SUBMITTED: 12 October 2021
PAPER ACCEPTED: 10 November 2021

The COST action CA16204 - Distant Reading for European Literary History issued several STSM (short term scientific mission) calls enabling action participants to visit organizations in other action member countries, with the goal to learn more about a topic in which a host organization is specialized, or to solve together a specific problem. The mission to procure comparative stylistic and morphosyntactic analysis of ELTeC texts using *stylo R* package[1] was organized within the third call (the first quarter of 2020), and in March 2020 I visited one of the institutions that participated in the creation of the package, the Polish language institute in Krakow.

The purpose of the short-term scientific mission was to perform experiments that would apply different levels of morphosyntactic annotations to a text from the ELTeC collection, in order to obtain accurate numeric comparisons between these different text incarnations. The main motivation was to test the usability of these morphosyntactic annotations in stylometric analysis and find the optimal combination of annotations that can provide better classification results in various scopes (gender, time period and authorship) related to ELTeC texts. If such approaches were proven to increase classification accuracy, at least to some extent, they could be applied for further analysis of texts written in languages with developed morphosyntactic annotation tools.

Work was done through several working sessions in which researchers involved in creation and maintenance of the *stylo* package participated. These included methodology overview, a workshop on classification using the *stylo R* package, discussion meetings, and one session devoted to interaction with everyday users. All tests were performed using the Serbian ELTeC texts. Balanced subsets – incarnations of the documents - were prepared for the experiments by cross sectioning two groups of variants: one was using the metadata

[1] Stylometry with *stylo*

(author, authors gender and time-period) and the other was using different levels of morphosyntactic annotation (part-of-speech – POS, grained part-of-speech, lemma, and word forms, that is, words with no-annotation) totalling in twelve different incarnations. These were all tested on classification reliability to establish their efficiency and create a comparative analysis. The idea was to answer the following and similar questions: i) What defines an author's style? Is it the use of word forms or the use of lemmas?; ii) Do different authors of different gender use different word combinations or word forms with different lemmas? iii) Is the use of a certain part of speech specific for a time period in literature?

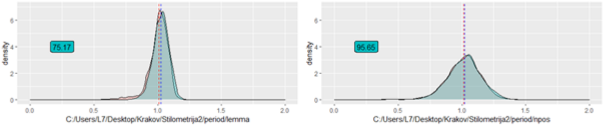


Figure 1. Delta distances distributions (same-class is blue and different-class is red) for the time period using lemmas (left) and for the gender using grained POS (right) outputted through R studio via R script prepared during the STSM.

The classification reliability was investigated by overlapping same-class and different-class cosine delta distance distribution. Overlapping areas of such distributions were procured as indicators of classification reliability for certain document incarnations. If same-class and different-class cosine delta distance distributions overlap, then accurate classification is improbable, and if, on the other hand, they do not overlap it indicates a higher probability of accurate classification using incarnations of that text, with the smaller area on the output graphs. In cases where distributions intersect multiple times, overlap or have close-by means, the classification reliability cannot be accurately measured; in that case, it is by default very low or none. For testing purposes an R script that uses the *stylo* package was compiled, which can be reused for testing the results for different classes, annotations and even languages, provided with the required text incarnations.

Classifications by gender and period were both found to be unobtainable for Serbian ELTeC texts. This was observed through multiple intersections of the series, close-by means and overall multitude of overlapped areas for these cases of simple binary classification. Two out of eight graphs for these classifications are shown as an example of the low reliability classification in Figure 1.

These results were explained by the narrow time window of the text origin for Serbian novels of ELTeC corpus, in regard to the period, and the lack of

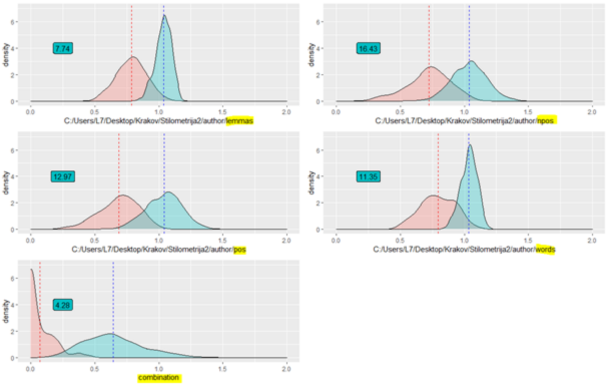


Figure 2. The comparison of delta distance distribution for different text incarnations regarding authorship attribution. Top left is for a text incarnation using only lemmas, top right is using only grained POS, middle left is using only POS, middle right is using an original text and bottom is using the combination of delta distances for all incarnations.

obvious differentiation of female and male writing styles in early 20 century Serbian novels (combined with the low representation of female authors in that period in general) regarding gender differentiation.

On the other hand, classification by authors (authorship attribution) yielded interesting results. The text collection comprised 80 novels written by 17 writers. Findings of previous stylometric research in Slavic languages were confirmed by this experiment. Classification using lemmatized text was found to be the most efficient, with a text incarnated into a series of part-of-speech tags (and using trigrams) falling behind the original text. The Grained POS (also using trigrams) was found to have the worst result in this scenario. The text incarnations with POS tags showed to have higher standard deviation. These results are presented in Figure 2.

However, the combination method (using distance matrices of different incarnations) developed on the last day of the STSM proved to be the most reliable, with up to two times better result than a lemmatized text, over a 17-class classification by authors, with possible error margin down to only 4.3% (see the bottom row of Figure 2). The obtained results prove that the STSM to Krakow was successful.

Acknowledgment

This STSM would not be possible without the hosts from the Polish language institute in Krakow – Maciej Eder and Joanna Byszuk, nor without the help of their keen colleagues from the Institute's Department of methodology.

Distant Reading Training School 2020: Named Entity Recognition & Geo-Tagging for Literary Analysis

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia

PAPER SUBMITTED: 2 November 2021
PAPER ACCEPTED: 15 November 2021

1 Introduction

Distant Reading Training School "Named Entity Recognition & Geo-Tagging for Literary Analysis" was organised virtually within the COST Action 16204: Distant Reading for European Literary History, on 22-25 March 2020. The host institution was the Faculty of Humanities and Social Sciences, University of Rijeka, Croatia, and it was organised within coordinated activity of WG2 "Methods and tools" and WG3 "Literary Theory and History".

WG 2 coordinates activities related to sharing, evaluating and improving methods and tools for distant reading research, with a focus (1) on tool and method adaptation and (2) on establishing best practices across Europe. Members of WG2 come from computational linguistics, text mining, computational stylistics, and digital literary studies. The work of WG3 concentrates on application of distant reading methods to literary history. Members of WG3 come from partners active in digital literary studies and mainstream literary history and theory.

Ten trainers from France, Italy, Norway, Poland, Croatia and Serbia introduced several topics related to Named Entity Recognition (NER) to the target audience, comprising researchers, especially early-career investigators (ECI), from participating countries interested in Distant Reading, Digital Literary Studies, Corpus and Computational Linguistics and/or Literary Theory and their methodological uses across national traditions. The training school included two workshops over the course of 3 days.[1]

[1] All training materials are available on Action's github, including: slides, notebooks and datasets.

2 Workshop 1: Introduction to Named Entity Recognition

The 2-day workshop introduced the task of Named Entity Recognition and described several annotation guidelines and campaigns. The practical part covered a) basic manual annotation with different tools (BRAT,[2] Inception[3] and Recogito),[4] and the analysis of disagreement between annotators, b) automatic annotation with easy-to-use tools such as CLARIN-PL NER tool suite[5] and NER&Beyond,[6] c) TEI-encoding of NER annotation, and d) practical exercises in analysing NE contexts as far as description, sentiment and perception are concerned. For practical reasons and better understanding of the procedures, the exercises were focused on English, but the workshop was addressed to speakers of all ELTeC languages, so that they could learn about NER to work on their collections. Therefore, examples from other languages were also presented.

Diana Santos from the University of Oslo gave an overview of the history of named entity recognition, starting with MUC(K) (1987-1998), IREX (1998), later ACE (2002-2008), CoNLL (2002; 2003), TimeML (2003), ENE (2004), HAREM (2006; 2008), TempEval (2007; 2010; 2013) up to recent SHINRA (2020) and concluding with the references to the research literature about NER.

Carmen Brando from the School for Advanced Studies in the Social Sciences presented several named entity recognition systems through several topics: tool pipelines for linguistic analysis and NER systems; challenges and features for NER systems; types of NER systems; manual annotation and evaluation and training of NER systems; some available out-of-the-box NER systems and output NE annotation formats.

Francesca Frontini from the Institute for Computational Linguistics in Pisa presented the ELTeC NER annotation campaign (Frontini et al. 2020), starting with requirements set by WG3 and motivation for ad hoc annotations, continuing with annotation guidelines containing explanations of category annotations and the annotation procedure (nested annotations,

[2] Brat developer site, github repository: ; Jerteh node used for DR NER 9 language collection

[3] Inception

[4] Recogito

[5] CLARIN-PL

[6] NER&Beyond

inclusion of determiners, ...), and finishing with the process wrap-up and alignment of annotations for various languages.

Ranka Stanković from the University of Belgrade presented the software infrastructure with tools related to NER: BRAT (Stenetorp et al. 2012) for manual annotation and NER&Beyond for formats and transformations. The annotation campaign encompassed the dataset preparation for all languages, dataset publishing (txt+ann), manual annotation or correction of automatic annotations and detailed and simplified annotation. A small experiment with inter-annotator agreement was conducted, to find out where the differences stem from and how to minimise them. Comparison of manual and automatic annotations indicated some problems, testing options and comparison issues (Šandrih Todorović et al. 2021).

Ioana Galleron from the Sorbonne-Nouvelle University and Carmen Brando focused on "translating" the results into TEI[7] (Text Encoding Initiative) annotation, trying to explain what TEI tags for named entities are, how to use them, for simple to more elaborated annotations, and how to convert txt files into TEI/XML file.

Ranka Stanković explained the annotation campaigns (Stanković et al. 2019) and practical work with BRAT, while Maciej Piasecki and Tomasz Walkowiak from the Wrocław University of Technology demonstrated Clarin tools for recognizing named entities and temporal expressions in Polish, English and German.

Carmen Brando gave an introduction to place-based analysis of literary texts: concepts and related work in spatial humanities.

3 Workshop 2: Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Google Maps API

Benedikt Perak from the University of Rijeka introduced Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Google Maps API. Within the *Data analysis task*, using the Google Colab platform and Python scripts, the geo-tagged data was loaded and converted to a Pandas dataframe object as a useful format for creating simple exploratory statistics, e.g. calculating the proportion of the geo-tagged data per language, per book, per period, etc. The *Representation of the*

[7] Text Encoding Initiative

geo-entities comprised two parts: getting the geo-coordinates and mapping geo-names as markers.

Getting the geo-coordinates (longitude and latitude of a place) is a necessary task for the geo-name representation on the map. For finding appropriate geo-data two methods were explored: Google Places API and Wikipedia Python package to explore Wikipedia data on geolocated entities.

The advantage of using the Google Places API[8] to find the geo-coordinates of the geo-name is the possibility to tap into vast information of the Google Places API, which returns information about a variety of categories, places, establishments, prominent points of interest, and geographic locations. One can search for places either by proximity or by text strings, and the Place Search returns a list of places along with summary information about each of them; additional information is available via a Place Details query. The downside of this approach is the need to open an account for this type of query system, which is free for 0–100,000 place requests per month. Using the Wikipedia Python package to explore Wikipedia data on geolocated entities makes accessing and parsing data from Wikipedia easy. The option to find Wikipedia entries by geo-names and geo-coordinates were explored, as well as the extraction of the additional data.

The *Folium Package* was used for mapping geo-names as markers and representing the data as markers with tooltip and HTML popup on the map. This representation helps literary scholars with the location of the narratives and the interpretation of literary texts.

The teaching materials included several Colab notebooks: NER Processing using NLP tools (spaCy), Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Python Client for Google Maps Services. The Jupyter Notebooks are available in the WG2_notebooks folder.

Acknowledgment

We are grateful to Antonija Primorac, a Working Group 3 Leader and Joanna Byszuk, a Working Group 2 Leader for organizing this school and to all the trainers for their valuables lectures.

[8] Places API

References

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020.*

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER& Beyond: The Archaic and the Modern Intertwined." In *Deep Learning Natural Language Processing Methods and Applications – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021),* edited by Galia Angelova, Maria Kunilovskaya, Ruslan Mitkov, and Ivelina Nikolova-Koleva, 1252–1260. INCOMA Ltd. ISBN: 978-954-452-072-4. https://doi.org/10.26615/978-954-452-072-4_141.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. "Named Entity Recognition for Distant Reading in Several European Literatures." *DH Budapest 2019.*

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "BRAT: a web-based tool for NLP-assisted text annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics,* 102–107.

Novels and Authors of the Serbian ELTeC Collection

Cvetana Krstev
cvetana@matf.bg.ac.rs
University of Belgrade
Faculty of Philology
Belgrade, Serbia

Ranka Stanković
ranka@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

1 Introduction

In the following subsections we will present the content of the Serbian subcollection in the European Literary Text Collection (ELTeC), both the basic collection and the extension.

The authors of articles in this issue of *Infotheca* used for reference the versions of Serbian ELTeC sub-collections dated October 20th, 2021. In the following subsections we will present the final version of the Serbian basic sub-collection, as well as the last version of the extension, and we will give notes about the differences between these two version.

2 List of novels Novels

In Table in this section we will give the list of all novels comprising the final version of SrpELTeC collection as well as the version the extended SrpELTeC collection dated December 2021.

- The following symbols are used to mark the *status* (**B**) of each novel:
  - **B** - a novel belongs to the basic set (SrpELTEC), and it was in this set at the time articles in this journal were written;
  - **B**$^+$ - a novel belongs to the basic set (SrpELTEC), and it was added to this set after the articles for this journal were written;

- **E** - a novel belongs to the extended set (SrpELTEC-ext), and it was in this set at the time articles in this journal were written;
- **E**$^{+}$ - a novel belongs to the extended set (SrpELTEC-ext), and it was added to this set after the articles for this journal were written;
- **B**$^{-}$**E**$^{+}$ - a novel was moved from the basic set (SrpELTEC) to the extended set (SrpELTeC-ext) after the articles for this journal were written;

.

– The *identifier* of each novel (**ID**) begins with SRP which is followed by a string given in the corresponding column. In the table it is linked to the Wikidata entry for the ELTeC edition of the novel.

– The *author's gender* (**AG**) њитх тњо поссибле values: M for men, F for women.

– The *novel's size* (**NS**) can have three values: S for short novels (less than 50,000 words long), M for medium sized novels (having more than 50,000 words and less than 100,000 words), L for long novels (having more than 100,000 words). The value S^{-} signifies that a text has less than 10,000 words and do not qualify as a "novel" (according to the Action's eligibility critera).

– The novel's *reprint count* (**RC**) can have two values: L for low (it was not reprinted in the period 1970-2010, or it was reprinted just once), H for high (it was reprinted more than once in the period 1970-2010).

– The *time slot* (**TS**) of the novel's first edition can have values: T1 for the period 1840-1959, T2 for the period 1860-1879, T3 for the period 1880-1899, and T4 for the period 1900-1920.

S	ID	Title/Author	Year	AG	NS	RC	TS
B	18520	*Два идола* Атанацковић, Богобој	1852	M	S	H	T1
B	18590	*Ђурађ Бранковић : историческиј роман* Игњатовић, Јаков	1859	M	M	L	T1
B	18620	*Једна женидба : слика из живота* Игњатовић, Јаков	1862	M	S	H	T2
B	18630	*Милан Наранџић* Игњатовић, Јаков	1863	M	M	H	T2
B	18631	*Кочина крајина : историјски роман* Ђорђевић, Владан	1863	M	M	L	T2
B^{+}	18680	*У гостионици код „Полу-звезде“ на имендан шантавог торбара : Приповетка из народног живота* Грчић, Јован Миленко	1868	M	S	H	T2
B	18690	*Гмунденско језеро : путничка новела* Ђорђевић, Владан	1869	M	S	L	T2
B	18691	*Какав је ко онако му и бива : приповетка из народна живота / написао за народ Панта Поповић* Поповић, Панта	1869	M	S	L	T2
B	18730	*Калуђер : истина и поезија* Суботић, Јован	1874	M	M	L	T2
B	18740	*Сељаци : приповетка из сеоског живота, из године 1857.* Јакшић, Ђура	1874	M	S	H	T2
B^{+}	18741	*Сирота Банаћанка* Јакшић, Ђура	1875	M	S	H	T2
B	18750	*Васа Решпект* Игњатовић, Јаков	1875	M	S	H	T2
B	18751	*Глава шећера* Глишић, Милован	1875	M	S	H	T2
B	18752	*Без оца и мајке* Радуловић, Пера	1875	M	S	L	T2
B	18760	*Чича Тима : приповетка из учитељског живота* Јакшић, Ђура	1876	M	S	L	T2
B	18780	*Пред зору : роман у два дела* Михајловић, Бранко [Љубиша Бранковић]	1878	M	M	L	T2
E^{+}	18781	*Потурченица Лејла : (црте из ратова за слободу)* Милићевић, Милан Ђ.	1879	M	S^{-}	L	T2
B	18790	*Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834* Милићевић, Милан Ђ.	1879	M	S	L	T2

S	ID	Title/Author	Year	AG	NS	RC	TS
B	18791	*Пастир краљ или Ослобођење Србије : историјска приповетка / написао К. Б.* Барунчић, Коста	1879	M	S	L	T2
B$^+$	18792	*Хајдуци : биљешке с пута по Рујну* Милићевић, Милан Ђ.	1879	M	S	L	T2
B$^+$	18793	*Љубав и патриотизам* Јеремић, Јован	1879	M	S	L	T2
B	18800	*Драгоцена огрлица: прича у своје време* Комарчић, Лазар	1880	M	M	L	T3
E	18801	*Школска икона у нашем селу* Лазаревић, Лаза К.	1880	M	S		T3
B$^-$E$^+$	18810	*Десет пара: прича из живота у вароши* Милићевић, Милан Ђ.	1881	M	S	L	T3
E	18811	*Вертер* Лазаревић, Лаза К.	1881	M	S		T3
E	18820	*Поета и адвокат : приповетка* Игњатовић, Јаков	1882	M	S		T3
B	18821	*Каваљер Лаза : карактерна црта из друштвеног живота* Ђорић, Никола В.	1882	M	S	L	T3
B	18840	*Из учитељичког живота* Гавриловић, Драга	1884	F	S	H	T3
B	18860	*Омер Челебија : приповијетка из живота српскога народа* Милићевић, Милан Ђ.	1886	M	S	L	T3
B	18870	*Бабадевојка* Гавриловић, Драга	1887	F	S	H	T3
B	18871	*Мој кочијаш : слике 1883 године* Комарчић, Лазар	1887	M	M	L	T3
B	18880	*Сањало* Поповић-Шапчанин, Милорад	1888	M	S	L	T3
B	18881	*Дух времена сад је таки!* Марковић Адамов, Павле	1888	M	S	L	T3
B	18882	*Шумарева ћерка : приповетка из српскога живота / написао С. В. Поповић* Поповић, Стеван В.	1888	M	M	L	T3
B$^+$	18883	*Патница : роман* Игњатовић, Јаков	1888	M	L	H	T3
B	18890	*Девојачки роман* Гавриловић, Драга	1889	F	M	H	T3

S	ID	Title/Author	Year	AG	NS	RC	TS
B	18891	*Борци : роман из сеоског живота. Свеска 1* Веселиновић, Јанко М.	1889	M	S	H	T3
B	18892	*Силазак с престола : роман / написао Карио Амурели* Тодоровић, Пера	1889	M	L	L	T3
B	18910	*Иконија везирова мајка: приповетка из XVII века* Мијатовић, Чедомиљ	1891	M	S	L	T3
B	18911	*Даница : роман из београдског живота* Јевтић, Стеван Ј.	1891	M	S	L	T3
B $^-$E$^+$	18920	*Рајко од Расине: приповетка с краја XVII века* Мијатовић, Чедомиљ	1892	M	S	L	T3
B	18921	*Србин и Хрватица или љубав и народност : роман : Из народносних размирица Срба и Хрвата у Загребу* Рогић, Душан	1892	M	S	L	T3
B	18922	*Бакоња фра-Брне : његово ђаковање и постриг* Матавуљ, Симо	1892	M	M	H	T3
B	18923	*Београдске тајне : исторички роман из српске прошлости, с краја прошлог века!. Св. 1* Тодоровић, Пера	1892	M	L	L	T3
B	18930	*Прве жртве: приповетка из српске прошлости* Гавриловић, Андра	1893	M	S	L	T3
E	18931	*Божићна печеница : приповетка* Сремац, Стеван	1893	M	S	H	T3
B	18932	*Сељанка : приповетка из сеоског живота* Веселиновић, Јанко М.	1888	M	S	H	T3
B	18933	*Смрт Карађорђева : исторички роман из недавне прошлости* Тодоровић, Пера	1893	M	M	H	T3
B	18934	*Један разорен ум* Комарчић, Лазар	1893	M	S	H	T3
B	18935	*Робињица Злата : приповетка из прошлости* Јовичић, Живојин	1893	M	S	L	T3
B	18940	*Радетића Мара: приповетка из сеоскога живота* Сретеновић, Михаило	1894	M	S	L	T3
B	18941	*Поп Ћира и поп Спира : приповетка* Сремац, Стеван	1894	M	M	H	T3
B	18942	*Синовац : оригинални роман* Радовић, Димитрије	1894	M	S	L	T3
B	18950	*Ивкова слава : приповетка* Сремац, Стеван	1895	M	S	H	T3

S	ID	Title/Author	Year	AG	NS	RC	TS
B	18951	*Кажњено неверство : најновији роман*					
		Светолик, Владимир	1895	M	M	L	T3
B	18960	*Господа сељаци: приповетка*					
		Костић, Тадија П.	1896	M	S	L	T3
B	18961	*Неједнака браћа : приповетка за народ*					
		Мамузић, Стеван	1896	M	S	L	T3
B	18962	*Конац дело краси : српској младежи приповеда Душан Ђурић*					
		Ђурић, Душан	1896	M	S	L	T3
B	18963	*Хајдук Станко : историјски роман*					
		Веселиновић, Јанко М.	1896	M	M	H	T3
B	18964	*Деспотова властела : роман из српске прошлости*					
		Гавриловић, Андра	1896	M	M	L	T3
B	18965	*Неимари : роман из новије српске историје*					
		Петровић, Коста	1896	M	M	L	T3
B	18966	*Назарени : роман*					
		Томић, Јаша	1896	M	M	L	T3
B	18970	*Грофица Агнеша Јанковић : роман*					
		Матијашић, Стеван М.	1897	M	M	L	T3
B	18971	*Горски цар : роман*					
		Ранковић, Светолик	1897	M	M	H	T3
B	18980	*Швабица*					
		Лазаревић, Лаза К.	1898	M	S	H	T3
B	18990	*Кнез Градоје од Орлова Града : приповетка из времена боја на Косову*					
		Мијатовић, Чедомиљ	1899	M	S	L	T3
B^{-}E^{+}	18991	*Увела ружа*					
		Станковић, Борисав	1899	M	S	H	T3
B	18992	*Лутајуће душе : приповетка за поуку : намењено задругарству*					
		Јањић, Велислав	1899	M	S	low	T3
B	18993	*Сеоска учитељица : роман*					
		Ранковић, Светолик	1899	M	M	H	T3
B	19000	*Порушени идеали : роман*					
		Ранковић, Светолик	1900	M	M	H	T4
B	19001	*Миланово школовање*					
		Тутуновић, Радојица В.	1900	M	M	L	T4
B	19002	*Крвави злочин у браку без љубави*					
		Кара-Радовановић, Павле	1900	M	S	L	T4

S	ID	Title/Author	Year	AG	NS	RC	TS
B	19010	*Гила : новела из сеоског живота* Поповић, Тодор Љ.	1901	M	S	L	T4
B	19011	*Пpичек Цара Душана у Дубровнику год. 1349. : историчка новела из XIVв.* Вулетић-Вукасовић, Вид	1901	M	S	L	T4
B	19012	*Ђул-Марикина прикажњa : приповетка* Димитријевић, Јелена	1901	F	S	L	T4
E	19020	*Једна угашена звезда : илустровани роман* Комарчић, Лазар	1902	M	S		T4
B	19021	*Општинско дете: роман једног одојчета* Нушић, Бранислав	1902	M	M	H	T4
B^{-}E^{+}	19022	*Покојникова жена* Станковић, Борисав	1902	M	S	H	T4
B	19023	*Славко* Ранковић, Драгутин Ј.	1902	M	S	L	T4
B	19024	*Харамбаша Мицко : српски војвода и заточеник у Фесану* Тасић, Димитрије С.	1902	M	M	L	T4
B	19025	*Страдија* Домановић, Радоје М.	1902	M	S	H	T4
B	19030	*Прво весеље : приповетка* Костић, Тадија П.	1903	M	S	L	T4
B	19031	*Две сестре или Самоубиство једне шваље : слика из београдског живота* Савић, Божа	1903	M	M	L	T4
B	19040	*Хаџи-Ђера : приповетка* Илић, Драгутин Ј.	1897	M	M	H	T3
B	19041	*За крухом* Ћипико, Иво	1904	M	M	H	T4
E	19050	*Просиоци : роман* Комарчић, Лазар	1905	M	S	L	T4
B	19051	*Женидба Пере Карантана* Ћоровић, Светозар	1905	M	S	L	T4
B	19060	*Новац : роман из београдског живота* Талетов, Пера С.	1906	M	M	L	T4
B	19061	*Хаџи-Диша : роман из живота старог Београда* Илић, Драгутин Ј.	1906	M	M	H	T4

S	ID	Title/Author	Year	AG	NS	RC	TS
B	19062	*Злочин једне свекрве : криминална приповетка : из скоре прошлости / по аутентичним подацима и судским актима написао К. Д. Јездић* Јездић, Коста Д.	1906	M	S	L	T4
B^{+}	19063	*Из земље плача* Тунгуз-Перовић, Радован	1906	M	M	L	T4
B	19070	*Фати-султан* Димитријевић, Јелена	1903	F	S	L	T4
B	19071	*Зона Замфирова : приповетка* Сремац, Стеван	1907	M	M	H	T4
B	19080	*Модерно робље : роман из живота босанских Срба* Петровић, Бошко Ст.	1908	M	M	L	T4
B	19090	*Пауци* Ћипико, Иво	1909	M	S	H	T4
B	19091	*Препорођај* Динић, Сретен	1909	M	S	L	T4
B	19100	*Дошљаци : роман* Ускоковић, Милутин	1910	M	M	H	T4
B	19101	*Нечиста крв* Станковић, Борисав	1910	M	M	H	T4
B	19102	*Радиша или какав нам учитељ треба на селу* Миодраговић, Јован	1910	M	L	L	T4
B^{-}E^{+}	19110	*Потрошене речи* Ускоковић, Милутин	1911	M	S	L	T4
B	19120	*Нове : роман* Димитријевић, Јелена	1912	F	M	L	T4
B	19121	*Беспуће* Милићевић, Вељко М.	1912	M	S	H	T4
B	19130	*Калуђер и хајдук : приповетка о последњим данима Србије у XV веку* Новаковић, Стојан	1913	M	M	high	T4
B	19131	*У фронт : приповетка из живота једног бившег краља* Ђорђевић, Владан	1913	M	S	L	T4
B	19132	*Јарани : приповетка* Ћоровић, Светозар	1913	M	S	L	T4
B	19140	*Чедомир Илић : роман* Ускоковић, Милутин	1914	M	M	H	T4
B	19141	*Кад шуме таласи* Ђуричић, Младен Ст.	1914	M	S	L	T4

S	ID	Title/Author	Year	AG	NS	RC	TS
B	19180	*Пре среће*					
		Јанковић, Милица	1918	F	S	L	T4
B^+	19181	*Војник Стојан : недовршен ратни роман*					
		Петровић, Драгомир С.	1918	M	M	L	T4
E	1918a	*Американка*					
		Димитријевић, Јелена	1918	F	S^-		T4
B	19190	*Ђакон Богородичине цркве*					
		Секулић, Исидора	1919	F	M	H	T4
E	19191	*Брђани*					
		Ћоровић, Светозар	1919	M	S		T4
B	19192	*У ћелијама*					
		Ћоровић, Светозар	1919	M	S	L	T4
B	19200	*Један од многих : роман из престоничког живота*					
		Шишковић, Драгомир	1920	M	S	L	T4
B	19201	*Деветсто петнаеста : трагедија једног народа*					
		Нушић, Бранислав	1920	M	L	H	T4
E	19202	*Соња*					
		Николић, Милан М.	1920	M	S	L	T4
B	19203	*Легија смрти : роман из Балканског рата 1912/1913.*					
		(написао Бранислав Јуришић)					
		Суботић, Каменко	1920	M	L	L	T4
E	1920a	*Пут Алије Ђерзелеза*					
		Андрић, Иво	1920	M	S^-		T4

3 List of Authors

In this section the list of authors is given with all their novels represented in either SrpELTeC or its extension. Each author's name is linked to its entry in Wikidata, while its VIAF number (if it exists) is linked to its record in VIAF database.

One note about the authorship: The author Pera Todorović is represented in SrpELTeC with three novels. One of these novels, *Београдске тајне : историски роман из српске прошлости, с краја прошлог века!* (Belgrade secrets: a historical novel from the Serbian past, the end of previous century) was until recently treated as written by an unknown author, since neither the book copy nor the library catalogue provided any data about the authorship. However, the further research of different sources proved that it was written by Pera Todorović.

ID	Author	Birth/Death	viaf/Year
	Андрић, Иво	1892–1975	97177322
1920a	*Пут Алије Ђерзелеза*		(1920)
	Атанацковић, Богобој	1826–1858	8970679
18520	*Два идола*		(1852)
	Барунчић, Коста		
18791	*Пастир краљ или Ослобођење Србије : историјска приповетка / написао К. Б.*		(1879)
	Веселиновић, Јанко М.	1862-1905	66316956
18932	*Сељанка : приповетка из сеоског живота*		(1884)
18891	*Борци : роман из сеоског живота. Свеска 1*		(1889)
18963	*Хајдук Станко : историјски роман*		(1896)
	Вулетић-Вукасовић, Вид	1853–1933	303498941
19011	*Причек Цара Душана у Дубровнику год. 1349. : историчка новела из XIVв.*		(1901)
	Гавриловић, Андра	1864–1929	14822825
18930	*Прве жртве: приповетка из српске прошлости*		(1893)
18964	*Деспотова властела : роман из српске прошлости*		(1896)
	Гавриловић, Драга	1854–1917	17281658
18840	*Из учитељичког живота*		(1884)
18870	*Бабадевојка*		(1887)
18890	*Девојачки роман*		(1889)
	Глишић, Милован	1847–1908	73846855
18751	*Глава шећера*		(1875)
	Грчић, Јован Миленко	1846–1875	37742606
18680	*У гостионици код „Полу-звезде“ на имендан шантавог торбара : Приповетка из народног живота*		(1868)
	Димитријевић, Јелена	1862–1945	70015495
19012	*Ђул-Марикина прикажња : приповетка*		(1901)
19070	*Фати-султан*		(1903)
19120	*Нове : роман*		(1912)
1918a	*Американка*		(1918)
	Динић, Сретен	1875–1949	55248036
19091	*Препорођај*		(1909)
	Домановић, Радоје М.	1873–1908	7623303
19025	*Страдија*		(1902)
	Ђорђевић, Владан	1844–1930	106964101
18631	*Кочина крајина : историјски роман*		(1863)
18690	*Гмунденско језеро : путничка новела*		(1869)
19131	*У фронт : приповетка из живота једног бившег краља*		(1913)

ID	Author	Birth/Death	viaf/Year
	Ђорић, Никола В.	1859–1913	22163155990601312095
18821	*Каваљер Лаза : карактерна црта из друштвеног живота*		(1882)
	Ђурић, Душан		
18962	*Конац дело краси : српској младежи приповеда Душан Ђурић*		(1896)
	Ђуричић, Младен Ст.	1889–1987	38235595
19141	*Кад шуме таласи*		(1914)
	Игњатовић, Јаков	1822–1889	4995002
18590	*Ђурађ Бранковић : историческиј роман*		(1859)
18620	*Једна женидба : слика из живота*		(1862)
18630	*Милан Наранџић*		(1863)
18750	*Васа Решпект*		(1875)
18820	*Поета и адвокат : приповетка*		(1882)
18883	*Патница : роман*		(1888)
	Илић, Драгутин Ј.	1858–1926	79412565
19040	*Хаџи-Ђера : приповетка*		(1897)
19061	*Хаџи-Диша : роман из живота старог Београда*		(1906)
	Јакшић, Ђура	1832–1878	47570198
18740	*Сељаци : приповетка из сеоског живота, из године 1857.*		(1874)
18741	*Сирота Банаћанка*		(1875)
18760	*Чича Тима : приповетка из учитељског живота*		(1876)
	Јанковић, Милица	1881–1939	41876311
19180	*Пре среће*		(1918)
	Јањић, Велислав		
18992	*Лутајуће душе : приповетка за поуку : намењено задругарству*		(1899)
	Јевтић, Стеван Ј.	1854–1904	68773127
18911	*Даница : роман из београдског живота*		(1891)
	Јездић, Коста Д.	1853–1930	58399970
19062	*Злочин једне свекрве : криминална приповетка : из скоре прошлости / по аутентичним подацима и судским актима написао К. Д. Јездић*		(1906)
	Јеремић, Јован	1860–1931	235008989
18793	*Љубав и патриотизам*		1879
	Јовичић, Живојин	1837–1908	
18935	*Робињица Злата : приповетка из прошлости*		(1893)
	Кара-Радовановић, Павле		
19002	*Крвави злочин у браку без љубави*		

ID	Author	Birth/Death	viaf/Year
	Комарчић, Лазар	1833–1909	14831335
18800	*Драгоцена огрлица: прича у своје време*		(1880)
18871	*Мој кочијаш : слике 1883 године*		(1887)
18934	*Један разорен ум*		(1893)
19020	*Једна угашена звезда : илустровани роман*		(1902)
19050	*Просиоци : роман*		(1905)
	Костић, Тадија П.	1863–1927	35277007
18960	*Господа сељаци: приповетка*		(1896)
19030	*Прво весеље : приповетка*		(1903)
	Лазаревић, Лаза К.	1851–1891	17270562
18801	*Школска икона у нашем селу*		(1880)
18811	*Вертер*		(1881)
18980	*Швабица*		(1898)
	Мамузић, Стеван		128149841986602842118
18961	*Неједнака браћа : приповетка за народ*		(1896)
	Марковић-Адамов, Павле	1855–1907	63747033
18881	*Дух времена сад је таки!*		(1888)
	Матавуљ, Симо	1852–1908	22201524
18922	*Бакоња фра-Брне : његово ђаковање и постриг*		(1892)
	Матијашић, Стеван М.		
18970	*Грофица Агнеша Јанковић : роман*		(1897)
	Мијатовић, Чедомиљ	1842–1932	42116339
18910	*Иконија везирова мајка: приповетка из XVII века*		(1891)
18920	*Рајко од Расине: приповетка с краја XVII века*		(1892)
18990	*Кнез Градоје од Орлова Града : приповетка из времена боја на Косову*		(1899)
	Милићевић, Вељко М.	1886–1929	81151776825018012654
19121	*Беспуће*		(1912)
	Милићевић, Милан Ђ.	1831–1908	64038897
18781	*Потурченица Лејла : (црте из ратова за слободу)*		(1879)
18790	*Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834*		(1879)
18792	*Хајдуци : биљешке с пута по Рујну*		(1879)
18810	*Десет пара: прича из живота у вароши*		(1881)
18860	*Омер Челебија : приповијетка из живота српскога народа*		(1886)
	Миодраговић, Јован	1854–1926	37838247
19102	*Радиша или какав нам учитељ треба на селу*		(1910)

ID	Author	Birth/Death	viaf/Year
	Михајловић, Бранко [Љубиша Бранковић]	1857–1918	85154625
18780	*Пред зору : роман у два дела*		(1878)
	Николић, Милан М.		
19202	*Соња*		(1920)
	Новаковић, Стојан	1842–1915	64052399
19130	*Калуђер и хајдук : приповетка о последњим данима Србије у XV веку*		(1913) (1913)
	Нушић, Бранислав	1864–1938	73889774
19021	*Општинско дете: роман једног одојчета*		(1902)
19201	*Деветсто петнаеста : трагедија једног народа*		(1920)
	Петровић, Бошко Ст.	1869–1913	
19080	*Модерно робље : роман из живота босанских Срба*		(1908)
	Петровић, Драгомир С.		1585145856998122920810
19181	*Војник Стојан : недовршен ратни роман*		(1918)
	Петровић, Коста	1858–1928	74867281
18965	*Неимари : роман из новије српске историје*		(1896)
	Поповић, Панта	1843–1918	
18691	*Какав је ко онако му и бива : приповетка из народна живота /написао за народ Панта Поповић*		(1869)
	Поповић, Стеван В.	1845–1918	262110860
18882	*Шумарева ћерка : приповетка из српскога живота / написао С. В. Поповић*		(1888)
	Поповић, Тодор Ђ.	1870–1957	1325145856883922920245
19010	*Гила : новела из сеоског живота*		(1901)
	Поповић Шапчанин, Милорад	1847–1895	79455054
18880	*Сањало*		(1888)
	Радовић, Димитрије		
18942	*Синовац : оригинални роман*		(1894)
	Радуловић, Пера		
18752	*Без оца и мајке*		(1875)
	Ранковић, Драгутин Ј.	1882–1956	252838823
19023	*Славко*		(1902)
	Ранковић, Светолик	1863–1899	29567902
18971	*Горски цар : роман*		(1897)
18993	*Сеоска учитељица : роман*		(1899)
19000	*Порушени идеали : роман*		(1900)

ID	Author	Birth/Death	viaf/Year
	Рогић, Душан	1855-????	94150747024316301259
18921	*Србин и Хрватица или љубав и народност : роман : Из народносних размирица Срба и Хрвата у Загребу*		(1892)
	Савић, Божа	1862–1927	305714596
19031	*Две сестре или Самоубиство једне шваље : слика из београдског живота*		(1903) (1903)
	Светолик, Владимир		
18951	*Кажњено неверство : најновији роман*		(1985)
	Секулић, Исидора	1877–1958	4943681
19190	*Ђакон Богородичине цркве*		(1919)
	Сремац, Стеван	1855–1906	66526515
18941	*Поп Ћира и поп Спира : приповетка*		(1894)
18950	*Ивкова слава : приповетка*		(1895)
19071	*Зона Замфирова : приповетка*		(1907)
18931	*Божићна печеница : приповетка*		(1893)
	Сретеновић, Михаило	1866–1934	305714513
18940	*Радетића Мара: приповетка из сеоскога живота*		(1894)
	Станковић, Борисав	1876–1927	76323147
18991	*Увела ружа*		(1899)
19022	*Покојникова жена*		(1902)
19101	*Нечиста крв*		(1910)
	Суботић, Јован	1817–1886	55020645
18730	*Калуђер : истина и поезија*		(1874)
	Суботић, Каменко	1870–1932	184145541835596601348
19203	*Легија смрти : роман из Балканског рата 1912/1913. (написао Бранислав Јуришић)*		(1920)
	Талетов, Пера С.	1875–1955	26486103
19060	*Новац : роман из београдског живота*		(1906)
	Тасић, Димитрије С.		305715401
19024	*Харамбаша Мицко : српски војвода и заточеник у Фесану*		(1902)
	Тодоровић, Пера	1852–1907	2788090
18892	*Силазак с престола : роман / написао Карио Амурели*		(1889)
18923	*Београдске тајне : историски роман из српске прошлости, с краја прошлог века!. Св. 1*		(1892)
18933	*Смрт Карађорђева : историски роман из недавне прошлости*		(1893)
	Томић, Јаша	1856–1922	5742092
18966	*Назарени : роман*		(1896)

ID	Author	Birth/Death	viaf/Year
	Тунгуз-Перовић, Радован	1879–1944	305581066
19063	*Из земље плача*		(1906)
	Тутуновић, Радојица В.		
19001	*Миланово школовање*		(1900)
	Ћипико, Иво	1869–1923	22935492
19041	*За крухом*		(1904)
19090	*Пауци*		(1909)
	Ћоровић, Светозар	1875–1919	136335
19051	*Женидба Пере Карантана*		(1905)
19132	*Јарани : приповетка*		(1913)
19191	*Брђани*		(1919)
19192	*У ћелијама*		(1919)
	Ускоковић, Милутин	1884–1915	14831361
19100	*Дошљаци : роман*		(1910)
19110	*Потрошене речи*		(1911)
19140	*Чедомир Илић : роман*		(1914)
	Шишковић, Драгомир		
19200	*Један од многих : роман из престоничког живота*		(1920)

Ideas and Observations from the Time of the ELTeC Corpus – a Selection of Quatations

Selection:
Cvetana Krstev

Panta Popović *Kakav je ko onako mu i biva : pripovetka iz narodna života* (SRP18691: 116)
(*the mother to her daughter on her wedding day*)

> Драгиња бризну сада још већма у плач, ал сада је и мати тешити стаде: Неплачи ћери моја, неплачи добре дете моје. Добром су детету свуда милостиви родитељи. Бог ти је дао здравља, а ја сам те научила раду. Моли се Богу дете моје ал и ради, кад дођеш у свекрову кућу, последња лези, а прва устани. Свекра и свекрву навек за савет питај; поштуј старије, и тебе ће млађи твоји. Спрам млађи ко спрам старији буди лепорека. Лепа реч гвоздена врата отвара. Навек нек си весела и растрешена; весело лице звери питоми, весело лице, најслађа је понуда на трпези. Никад се немој мргодити. Мргодење здраво гадно на лепом лицу стоји. Нек ти се навек умиљатост по лицу разлева, ко рујна зора по тијом небу. Тако ради ћери моја, па ћеш видити да умиљато јагње две овце сиса.

Milan Đ. Milićević *Hajduci : bilješke s puta po Rujnu* (SRP18792: 138–139)
(*the writer*)

> Од неког времена почели су наши љекари упућивати по кога болесника да проведе љети у Златибору по неколике надјеље у борју и јељацима. Ако је вјеровати оном што се прича, мало се ко вратио отуда, а да се није хвалио но повратку. А док се некад забјелуцају на Чиготи, на Груди, по Торнику, на Смиљанића Закосу, или на Цареву Пољу, чисте гостионице за љетње путнике, колико ће њих, сједећи на тијем висинама, гутати очима природне љепоте које су се разастрле на сваку страну! На исток одатле виде се, као нешто срдити, обли Овчар и оцијепљени Каблар, можда зато, што их је раздвојила бујна Морава; на сјевер ћуте већ за облаке припојени: Маљен, Букови,

> Таор, и Јабланик; на западу су понајближа брда: Груда, Шарган, Тара, и Звијезда а на југу је тек право чудо; иза дробнијех најближих висова босанскијех и херцеговачкијех, којима се не да ни број ухватити, ни облик описати, помаља се снијегом покривена Љубишња, и румене се према јутарњем сунцу дивотни Комови и Дурмитор.
> Кад очи све то сагледају, онда душа осјећа потребу да се диви нечему незнаному, нечему бескрајному, нечему — што срце умије да осјећа, а језик не умије да искаже...
> У средини овакијех позорја, човек може да пожели и оно што не само не може да буде него може и смијех да изазове. Хтјело би му се да се дигне из гроба послије једно сто двјеста година мирна живота и разумна рада! Бар би му имале очи на што да погледају!...

Jovan Jeremić *Ljubav i patriotizam* (SRP18793: 68–69)
(*the writer*)

> «Нема опасније звери од жене!»
> Зашто нисте лепе душе, као што сте лепа лица? Ох, да сте такове, не би Бандино говорио онако гњевно о вама, не би се морао онако љуто огрешити... Па шта је рекао? Да л’ смем, да л’ могу то поновити? Грозне и крвавим гњевом заливене су то речи... „Аух женскињо! ма да потресам кости покојне ми матере проклињући ти род.... проклете да сте, колико вас је на земљи анђеоског лица а демонског срца!“...[1]
> Не срдите се, миле сестре, ја ово већ не би рекао, ја сам са великим страхом и поновио те туђе речи; а и Бандино се морао пренаглити кад је проклео све женске... гњев га савладао, заборавио се у грозном сећању на неку женску.... Да љуте клетве. Нисте ви сва за клетву не, некоје сте и за штовање, премда ретке.... и генији су ретки?

[1] F. D. Gveraci (authoor’s note)

Jakov Ignjatović *Trpen spasen : roman u tri knjige* (iz 1897:112)[2]
(*the writer*)

> Глађеновић је знао нешто талијански, научио је у Земуну од једног келнера Талијанца. Ослови Павана, Паван је штрбецао нешто немачки; што овај не зна немачки, то овај дода талијански. То је обојици мило било; а и самој Морлакињи је мило што је знала са Глађеновићем разговарати, и то српски ил' хрватски, ил' приморски, ил' нашки, јер је то свеједно.

Pera Todorović *Silazak s prestola : roman* (SRP18892: 568)
(*Ratko, the minister of interior affairs*)

> *Граби на све стране власт, и што једном дочепаш, чувај као очи у глави. Од сеоског пудара па до министра свако место има свој значај и своју вредност, ни једно није за презирање и све их треба грабити себи.* Власт, власт и опет власт! Ето у ово неколико речи састоји се цео програм наш. Власт кроз сва времена, власт под свима министрима, <pb n="569"/> власт у свима околностима, власт дању, власт ноћу — ето ту је све.

Pera Todorović *Beogradske tajne : istorijski roman iz srpske prošlosti, s kraja prošlog veka* (SRP18923: 484–489)
(*Black Warrior to Belgrade vizier*)

> Црни Ратник бејаше се одушевио и распалио; план за планом, и слика за сликом низале су се из његова одушевљена, вешта разлагања и везир је осећао како овај топли, срдачни говор и њега све више и више загрева и осваја.
> Међу тиа раздрагана душа прекаљена борца-родољуба узлетала је све на веће и веће висине.
> Како би дивно могло процветати Балканско полуострво кад би, у братској слози, господарили њиме само његови народи!
> Пре свега какав красни народносни мозаик! Грци, Арбанаси, Срби, Бугари, Румуни, европски Турци! Пет-шест разних народа, готово подједнако бројно јаких, поређали се један поред другога, упућени самом природом да један другом пруже руку и да се помажу.

[2] This novel has not been yet completely prepared and thus has not the fixed ID.

Па какав диван материјал људски, каква снажна раса! Телесно крепки, развијени, здрави и одарени чудном лепотом обличја и телесна строја; душевно млађани, ведри, бистри, одарени високим способностима ума и узвишеним полетима чисте душе своје. А уз то јуначни, ратоборни, племенити, родољубиви, готови на свакојаке жртве; најзад велики устаоци, вредни, жудни знања и науке, а способни за најозбиљније умне радове — ови народи лако би ми могли постати пионери културе и дика целокупна напретка људског, да цела Европа с поносом може погледати на Балканско тропоље...

[...]

Дакле ви збиља верујете да међу балканским народима може бити неке везе и заједнице и ту дајете места и Турцима, и ако су они дошљачка, господарећа класа? — упита везир. — Зашто не би веровали у тај савез! Та, забога, и саме се животиње удружују, кад нађу да им је то потребно и корисно? Па зашто не би могли паметни и свесни људи створити једну заједницу, кад су тако очевидне користи од ње, и кад је тако јасно да је то једини пут да буду слободни на дому своме и да живе мирно, као срећни и независни народи, који своју судбу држе у својој рођеној руци. Што рекосте о дошљацима, то је како се узме. Сви смо ми у неку руку дошљаци овде на балканском тропољу, само што су неки дошли раније, а неки опет доцније. Не знам за што би та разлика у времену, кад је ко дошао, давала једном већа а другоме мања права! Сви смо ми подједнако деца природе, а ова лепа земља наша је заједничка домовина, мила Турцима тако исто као и нама Хришћанима и народима другога порекла. Што се пак тиче господареће класе, ње бити не може, и наша света дружина за то је и постала, да уништи то господарење једне класе над другом. Турци који улазе у нашу дружину мудри су и увиђавни људи. Они виде да су Турци истина класа која данас влада, али они прозиру тако исто и то, да та владавина не може дуго трајати, и да наступају времена где би Турци од класе која влада, лако могли постати класа којом се влада. С тога је у њином рођеном интересу да тога господарења класе над класом у опште не буде, и место господара и робова да се утврди ред слободних грађана и слободне сложне браће.

Stevan Mamuzić *Nejdenaka braća : pripovetka za narod* (SRP18961: 43–44) (*Housekeeper Mladen*)

„Па ето то. Молим ја вас господо, и ви друга браћо, од кад се ова наша граница почела укидати, ево ти код нас свакојаког света. А ко ти ту није? Ту су ти Тотови, ту Швабе, ту Бачвани. Па ти тај свет купује, покупова све земље од нас, од наших људи, да ти је то страхота. То не ваља, то ја не би дозволио. Ја би сваком заповедио, да седи онде, где се родио. А какви су то послови, доћи у туђе село, па га прекупити. То није пре било. А што ви то не гледате, а господине?” упита газда Младен бележника.
Овај се смешио.
„Е, мој газда Младене, није то тако, ко што ти мислиш" рече бележник. „истина, да сте ви пре били дуго војници, одбијени од куће. Одбијали сте се тако и од рада. Него није само зато, него и за нешто друго. Дација била мала, дуван си сејао, ракију си пекао, вино буд зашто куповао. Ране ти је требало тек толико, да се прехраниш. Толико си и сејао. На више ти њива коров и трње расли. Зар није тако, а?“.

Stevan Matijašević *Grofica Agneša Janković* (SRP18970: 142–143) (*Pavle, the main character*)

Дођох у Париз. Да ли треба да вам опишем утисак ове промене? Како сам се осећао у Паризу? То је било неко чудновато помешано осећање дивљења и туге. Јер на мене није утицао само поглед на дивовске сразмере овога светскога града, већ и помисао на малене почетничке несређене прилике моје домовине. Ја сам у први мах био уништен, и сам се чудим како сам се подигао. — Станем на један угао и погледам у један фењер. На огромној гранитној плочи стоји диван гвоздени ступ са неколико кракова на сваком са пет светњака. Гледам га. Кад би се код нас овако што подигло у Крагујевцу, то би била читава парада, војска би маширала и банда би свирала. То би било свечаност.[3] Погледам на други угао: а оно тамо стоји исто тако чудовиште; па на трећем исто, на четвртом исто.

[3] Слично томе било је неколико десетина доцније у Београду, када се открио некакав водоскок од прста дебео који је после неколико дана усахнуо. (author’s note)

Boža Savić *Dve sestre ili samoubistvo jedne švalje : slika iz beogradskog života* (SRP19031: 57)
(*the editor of „Malog žurnala“*)

> — Онда ћемо покушати преко новина, приметиће уредник, али, право да вам кажем, ја сам приметио, да су владе баш оне награђивале, које су опозициони листови жигосали као неваљале људе. Изгледа ми, да оне нарочито таке ниткове и траже, јер су готови на све, што је неваљало, чак и на злочин. Часне и вредне чиновнике оне запостављају, гоне, премештају, многе пензионишу и отпуштају, а силеџије и несавесне нарочито одликују.

Boža Savić *Dve sestre ili samoubistvo jedne švalje : slika iz beogradskog života* (SRP19031: 238–239)
(*the writer*)

> Али људско друштво напредује. Развиће његово, у напреднијем и савршенијем правцу, никакве мере на свету не могу спречити. Свакако ће доћи време, када ће се у сваком људском бићу гледати човек, чија ће се права поштовати, а не, као до сад, немилице газити. Доћи ће време опште једнакости, какву су желели и данас желе најчаснији борци, који се у данашњем друштву тако гоне и кињe.

Vladan Đorđević *U front : pripovetka iz života jednog bivšeg kralja* (SRP19131: 94–95)
(*Mr. Baron*)

> — Ама да не претерујете по мало г. бароне? — упита Емилијан смешећи се.
> — Ни мало, Монсењеру. Земља је приморана да гради једну велику железницу немајући за то ни својих стручњака ни својих капитала. Она се морала задужити, а страни предузимачи зарадили су на томе послу сто на сто. Међутим цео тај грдан дуг наваљен је на ту државицу поглавито за угодност запада при његовим путовањима на исток, јер од земаљских производа могу да се извозе том железницом само они који се налазе близу и дуж те линије. Због тога та железница не зарађује ни толико колико износе њени велики режијски трошкови, и држава не само што нема никакве користи од железнице за коју се толико задужила, него мора још да додаје од

> својих прихода из других државних извора и тако је дошло до тога, да Силванија која нема више од 70 милиона државних прихода годишње, мора и од те суме да одваја 20 милиона на ануитете државног зајма. Кад би се извршило само оно што је у програму Г. Министра било предложено за поправку земљорадње, сточарства, шумарства, воћарства и винарства, удесетостручили би се земаљски производи Силваније, а кад би се извршила предложена мрежа друмова и железница, удесетостручио би се и њихов извоз. Поред тога, толика нова саобраћајна средства учинила би тек могућном експлоатацију нађених и откривање нових рудничких блага земаљских. Само то богаство руда, кад би се како треба експлоатисало, било би у стању за неколико деценија исплатити све државне дугове.</p>

Dragomir S. Petrović *Vojnik Stojan : nedovršen ratni roman* (SRP19181: 99–100)
(*Milojka's mother to Milojka and Stojan*)

> Бежите, јер Беганова мајка, кажу, да не плаче. Проћи ће ово као и све друго, као свака зараза, као скакавци, као лањски снег...

Author Guidelines

All *Infotheca* articles are published both in English and Serbian in the same issue. Authors should submit their articles in one of the languages; only after the notification of acceptance the translated article is expected (for Serbian authors; for all other authors translation from English to Serbian is provided by the journal). Except the printed edition, all articles are also published in the online edition in open access.

PAPER CATEGORIZATION

For documents accepted for publishing which are subject to review, the following categorization in the Journal applies:

1. Scientific papers:
   - Original scientific paper (containing previously unpublished results of authors' own research acquired using a scientific method);
   - Review paper (containing original, detailed and critical review of a research problem or a field in which authors' contribution can be demonstrated by self citation);
   - Preliminary communication (original scientific work in progress, shorter than a regular scientific paper);
   - Disquisition and reviews on a certain topic based on scientific argumentation.
2. Scientific articles presenting experiences useful for advancement of professional practice.
3. Informative articles can be:
   - Introductory notes and commentaries;
   - Book reviews, reviews of computer programs, data bases, standards etc.
   - Scientific event, jubilees.

Papers classified as scientific must receive at least two positive reviews. The opinions of the Editorial Committee do not have to correspond to those expressed in the published papers. Papers cannot be reprinted nor published under a similar title or in a changed form.

ELEMENTS OF MANUSCRIPTS

For scientific or professional papers the following data should be provided:

1. Papers should not normally exceed 15 A4 pages, Times New Roman 12pt. For longer articles the authors should contact the journal editors.
2. Names and surnames of all authors should be written in the sequence in which they will appear in a published paper.
3. After each author's full name, without titles and degrees, an e-mail address should be specified as well as the full and official name of his or her affiliation. (For large organizations full hierarchy of names should be specified, top down).
4. The submission date should be provided.
5. The authors should suggest the category of their paper but the Editor-in-Chief is responsible for the final categorization.
6. An informative abstract not normally EXCEEDING 200 WORDS that concisely outlines the substance of the paper, presents the goal of the work and applied methods and states its principal conclusion, should accompany the paper. The abstract should be supplied in both languages used for publication. In the abstract, authors should use the terms that, being standard, are often used for indexing and information retrieval.
7. Authors should supply at least 3 but not more than 10 keywords separated by commas that designate main concepts presented in the paper. The list of keywords should be supplied in both languages used for publication.
8. If paper derives from a Master's thesis or Doctoral dissertation authors should give the title of the thesis or dissertation, as well as a date of its submission and names of responsible institutions.
9. If the paper presents the results of authors' participation in some project or program, authors should acknowledge the institution that financed the project in a special section "Acknowledgment" at the end of the article, before the "Reference" section. The same section should contain acknowledgment to individuals who helped in the production of the paper.
10. If the paper was presented at a Conference but not published in its Proceedings, this should also be stated in a separate note.
11. Authors can use footnotes, while endnotes are prohibited; however, too long footnotes should be avoided. Authors can add appendices to their paper.
12. The referenced material should be listed in the section "References" at the end of the paper. In the reference list authors should include all information necessary for locating the referenced work. All items referenced

in the text should be listed here; nothing that was not referenced in the text should appear in this section.

EDITING CONVENTIONS FOR ACCEPTED PAPERS

1. Papers should be prepared and submitted using LATEX(the journal style and all packages can be downloaded from the journal web site). Authors that are not familiar with LATEXcan prepare their papers using Word, as .doc, .docx, .rtf or .txt documents. These authors should not use any special formatting – the final formatting and transformation to LATEXwill be done by the Infotheca team.
2. The papers written in Serbian should use CYRILLIC alphabet because they will be printed in that script. The only exceptions are those parts of the text for which the use of the other script, such as Latin, is more appropriate. All scripts should be represented using Unicode encoding, UTF-8 representation.
3. Title of the paper should not be written in capital letters. The authors should keep the length of titles reasonable – preferably less than 90 characters. For all titles authors should provide a shorter title that will be used for page headers.
4. Italic type may be used to emphasize words in running text, while bold type or italic bold type can be used if necessary. Underlined text should be avoided. Please do not highlight whole sentences or paragraphs.
5. Paper can be divided in sections and subsections, but more than two levels of the section headings should be avoided. All sections and subsections will appropriately numbered. Appendices, if any, should come at the end of the paper and they will also be appropriately labeled. If using lists, do not use more than two levels of nesting.
6. All paragraphs should be separated by one empty line (one Enter).
7. Authors should avoid too wide tables keeping in mind that the journal is published on A5 paper and. All tables, illustrations, diagrams and photographs should not be wider than 72.5 mm (the width of one column) or (exceptionally) 150 mm (the width of the page). All illustrations should be prepared in some lossless format, for instance .png, .tif or .jpg and their resolution should be at least 300 dpi.
8. The authors are kindly requested to add (if possible) the link to the screen from which a screenshot was taken. When taking a screen shot of a part of some screen, authors are advised to use the Zoom possibility of the browser or other program. For diagrams that are produced with Excel, please provide the original .xls document.

9. All tables, illustrations, diagrams and photographs should be prepared as separate files, both in black-and-white for printing and in color for the on-line version. Captions that should be below tables, illustrations, diagrams or photographs should remain in the text. Each file should have the same name as the file containing the main text, followed by the type of material to which the ordinal number in the text is added. For instance, the file containing the fourth figure of the paper "Example" should be named Example_figure_4.
10. Please add additional document(s) that explain some specific aspects of formatting required for your paper, for instance, formulas prepared in LaTeXin a .pdf format.
11. URL addresses that appear in the paper should be placed in footnotes; the date when the site was visited should be given.

REFERENCES AND CITATION

1. Referenced material should be listed at the end of the text, within the unnumbered section References. The reference section should be complete; references should not be omitted. This section should not contain any bibliographic information not referenced in the main text. Referenced items should not be mentioned in footnotes.
2. Entries in the reference list should be ordered alphabetically by authors or editors names, or publishing organizations (when no authors are identified). If this list contains several entries by the same authors, these entries should be ordered chronologically.
3. For preparation of a reference list use Chicago Manual of Style reference list entry (www.chicagomanualofstyle.org).
4. Full names of journals, and not their short titles or acronyms, should be specified. Use the 10-point type for entries in the reference list.
5. All authors, whether they prepare their articles using LaTeXor Word, will prepare all the items from their References section using BibTeX templates that are given for all the examples at the Infotheca web site (http://infoteka.bg.ac.rs/index.php/sr/upu-s-v-z-u-r).