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 Mamužić 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).

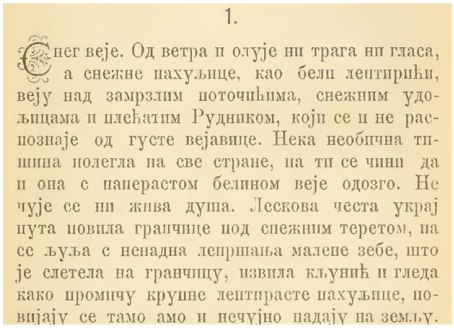
It was the responsibility of the Digitization Department of the University Library to scan the selected material. For these purposes, a Robotic book scanner,[12] type RBS 3.0 from 2014, was used. The output scans were in JPG format with a resolution of 300dpi. From the technical point of view, all settings, 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, вију 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 и о.тује
+++пи+++ни+++ии+++ трага
+++пи+++ни+++ии+++ гласа, а снежне
+++пахуљице+++, као +++бели+++
+++лептирићи+++, вију 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.

•J®.нег веје. Од ветра п о.тује нп трага нп
гласа, а снежне пахуљпце, као белп
лептирпћп, вију 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
спежннм теретом, па се љуља с ненадна
лепршања малене зебе, што је слетела на
гранчицу, пзвпла кљунић и гледа како
промичу крупне лептрасте пахуљпце,
по- впјају се тамо амо и нечујно падају на
зем.ву.

<p>Снег веје. Од ветра и олује
ни трага ни гласа, а снежне пахуљице,
као бели лептирићи, вију 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.

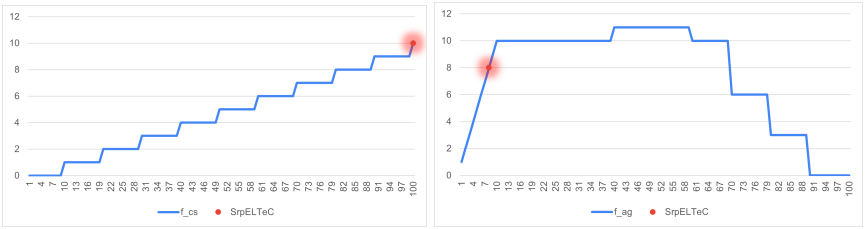


Figure 2. Collection size factor $f_{cs} = f(N_{novels})$ (left); Author gender factor $f_{ag} = f(AG)$ (right). Red dots are factor values for SrpELTeC.

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).

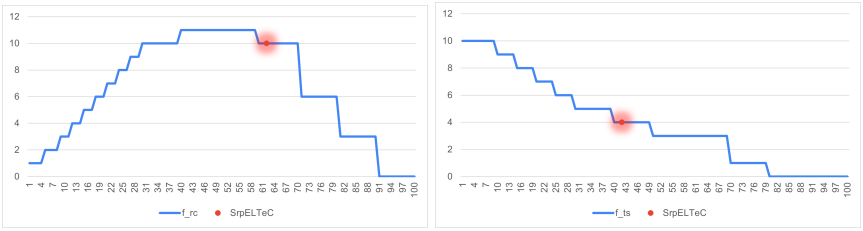


Figure 3. Reprint count factor $f_{rc} = f(RC)$ (left); Time slot factor $f_{ts} = f(TS)$ (right). Red dots are factor values for SrpELTeC.

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).

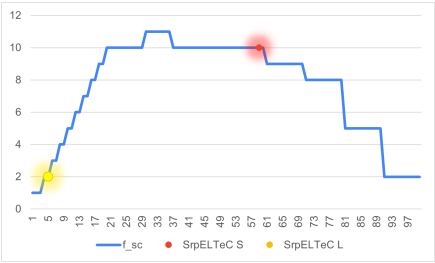


Figure 4. Size category factors: $f_{scs} = f(SCS)$ for short novels, $f_{scl} = f(SCL)$ for long novels. The red dot is the factor value for short novels, the yellow dot is the factor value for long novels for SrpELTeC.

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).

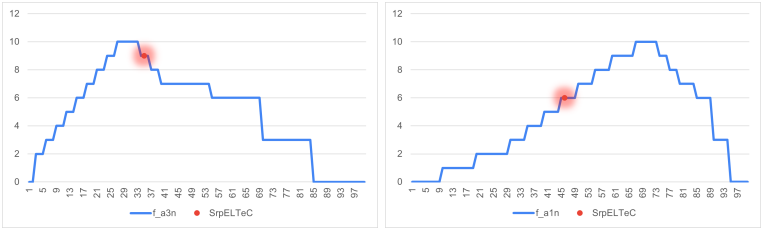


Figure 5. Three novels per author $f_{a3n} = f(A3N)$ (left); One novel per author $f_{a1n} = f(A1N)$ (right). Red dots are factor values for SrpELTeC.

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.

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

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

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

$E5C$ for SrpELTeC is:

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

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 повѣсть стъ сочиньна 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.

Тртовац, 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.