# SrpELTeC: A Serbian Literary Corpus for Distant Reading

Ranka Stanković, Cvetana Krstev, Duško Vitas




SrpELTeC: A Serbian Literary Corpus for Distant Reading | Ranka Stanković, Cvetana Krstev, Duško Vitas | Primerjalna književnost | 2024 | |

10.3986/pkn.v47.i2.03

http://dr.rgf.bg.ac.rs/s/repo/item/0008690

# SrpELTeC: A Serbian Literary Corpus for Distant Reading

Ranka Stanković, Cvetana Krstev, Duško Vitas

University of Belgrade, Faculty of Mining and Geology, Djusina 7, 11000 Belgrade, Serbia
https://orcid.org/0000-0001-5123-6273
ranka.stankovic@rgf.bg.ac.rs

Language Resources and Technologies Society, Studentski trg 3, 11000 Belgrade, Serbia
https://orcid.org/0000-0003-3328-9392
CvetanaJK@gmail.com

Language Resources and Technologies Society, Studentski trg 3, 11000 Belgrade, Serbia
https://orcid.org/0000-0003-4194-692X
dusko.vitas@gmail.com

*The article presents SrpELTeC, a corpus developed within the COST action Distant Reading for European Literary History (CA16204). All novels in SrpELTeC were selected, prepared, and annotated using the common principles established for all language collections in the European Literary Text Collection (ELTeC). The challenges and solutions in preparing SrpELTeC from scratch are outlined. All novels were manually encoded in TEI with rich metadata and structural annotation. The automatic annotation included POS-tagging, lemmatization, and named entities, relying on Natural Language Processing resources developed and maintained by the JeRTeh Language Resources and Technologies Society. The integration of SrpELTeC with Wikidata was supported with a set of SPARQL queries for the retrieval of metadata with different visualization options. Recent activities within the COST Action NexusLinguarum–European Network for Web-centred Linguistic Data Science (CA18209) are related to the linked data version of SrpELTeC using the NLP Interchange Format. All versions of SrpELTeC are freely available under the CC-BY license.*

Keywords: digital humanities / Serbian literature / text corpora / distant reading / linked data / named entity recognition / text analytics

## Introduction

The paradigm of distant reading involves the use of computer methods for the analysis of large collections of literary texts. The goal of these analyses is to complement the methods used in literary theory and history. Franco Moretti proposed reading methods that involve works outside the

established literary canon, which, following Margaret Cohen, he refers to as the "great unread" (Moretti 55). A novelty that Moretti suggests for the study of literature is the use of patterns, statistics, paratexts, and other properties that literary studies tended to disregard.

Applying methods of distant reading necessitates a careful selection of works based on firm criteria and an equally careful preparation of the selected works. The focus of the COST action Distant Reading for European Literary History (CA16204), which ran from 2017 to 2022, was the preparation of a multilingual resource dubbed the European Literary Text Collection, or ELTeC (Odebrecht et al.; see also Burnard et al.). The core of ELTeC contains a hundred novels first published between 1840 and 1920, covering twelve languages commonly spoken in Europe, each forming linguistic sub-collections (Schöch et al.). In addition, for nine languages partial collections of under a hundred novels were developed, while for six languages extended collections were developed as well.

The mandatory criteria for selection demanded that each work belong to narrative prose (as a novel or a long story), include a minimum of 10,000 words, and appear in first edition between 1840 and 1920. In order to be included in a certain language sub-collection, the work had to be originally written in that language, since translations were not foreseen. Preference was given to works that were published as books, rather than in installments in serial publications.

Additional conditions were set for the composition of each sub-collection, with the idea to ensure, on the one hand, the diversity of the works represented and, on the other hand, a comparative analysis of sub-collections and the application of key methods for the statistical analysis of texts. These additional criteria for the desirable corpus balance involved each sub-collection size, genders of authors, the lengths of texts and the number of their editions, even coverage of the period 1840–1920, and the number of novels per author. A sub-collection had to contain a hundred works that qualify as novels according to the mandatory criteria; optimally, it would also strike a balance between male and female authors. Canonical works as well as unknown and forgotten works were to be represented, where the number of editions of a work was used as a measure of its canonicity. The selected time period of first editions divided into four periods lasting 20 years each had to be evenly represented in each sub-collection, and each of these periods had to be represented by 20–25 works. According to their length, works were divided into short (10,000–50,000 words), medium-length (50,001–100,000), and long texts (more than 100,000 words). A sub-collection

was to contain at least 20% works of all lengths, ideally 30–40%. Additionally, a sub-collection had to contain nine to 11 authors represented by exactly three works (which would, for example, allow testing automatic authorship checking systems), while all other works had to be written by different authors to ensure a sufficient level of diversity.

The sub-collections for Czech, German, English, French, Swiss German, Hungarian, Polish, Portuguese, Romanian, Slovenian, Spanish, and Serbian are completed with 100 works. The additional nine languages, namely Greek, Irish, Croatian, Italian, Lithuanian, Latvian, Norwegian, Ukrainian, and Swedish, have incomplete sub-collections.

## The Serbian sub-collection of novels (SrpELTeC)

The Serbian sub-collection of novels, SrpELTeC, was created by a research team led by Cvetana Krstev. Given the aforementioned set selection criteria and the demand for balance, it comes as no surprise that creating a sub-collection of Serbian novels was not a trivial task, and that its development required much more effort than, say, in the case of English or French. First and foremost, prose writing in the Serbian language, and especially novel writing, appeared later than in most European countries, namely with the emergence of realism, which prevailed as a direction in the last three decades of the nineteenth century (Deretić, *Istorija* 362). This practically means that the set of works from which works could be chosen in order to better satisfy the balance criteria for Serbian was not as rich as for many other languages.

A number of Serbian literary works from this period, mostly canonical ones, were already digitized. Unfortunately, the way they were digitized did not allow us to include these digital editions in the SrpELTeC sub-collection. The library that contains the greatest number of Serbian literary works from the nineteenth and early twentieth centuries is called Antologija srpske književnosti (The Anthology of Serbian Literature) and was developed by the Faculty of Education at the University of Belgrade in cooperation with Microsoft. However, as these digital editions lack metadata, we cannot tell which editions of works included in this Anthology were used for digitization. It was therefore necessary to create the Serbian sub-collection from scratch. This took place in several steps, which will be briefly described below.

The first task was the compilation of a list of works that meet the selection criteria. The initial list with significant works and novelists was compiled from scholarly books by Jovan Deretić and Živan

Milisavac (Deretić, *Srpski*; Milisavac). The real challenge was to find so-called marginal works of Serbian literature, that is, books that were published only once or twice and whose authors are mostly forgotten today. The list was further enhanced with works that were found by searching the common catalogue of Serbian libraries COBISS+ by selecting appropriate values for the type of content or literary genre (novel, short story, short prose, etc.) and applying necessary restrictions on the language and the year of publication. Some works were retrieved from Belgrade dealers of antique books, while some suggestions were obtained from already acquired old books which contained appended lists of published works by the same publisher. Finally, a list of more than 150 candidates for the SrpELTeC sub-collection was obtained.

The next issue was to find and scan the books themselves, using the first editions whenever possible (Trtovac et al.). Books were mostly found in Belgrade national and university libraries and private libraries of project participants. Metadata were assigned to each novel of SrpELTeC, including data on their authors and publication, data on editions used for the collection, data concerning balance criteria, as well as data on institutions and individuals acknowledged for their help with the production of SrpELTeC (Krstev, "Serbian Part"). The metadata enables analyses of, for instance, titling practices of Serbian narrative literature in the years 1840–1920 (Patras et al.). In order to enable efficient metadata exploration, metadata are not only prepared as Text Encoding Iinitiative novel headers and Comma Separated Values files, but are also stored in Wikidata (Ikonić Nešić et al., "Serbian ELTeC"). This will be illustrated below.

## The digitization pipeline

The pipeline for the digitization process is presented in Figure 1. Scanning and optical character recognition of a chosen work (step 2) was followed by the manual correction and annotation (steps 3 and 4) in which a number of volunteers helped. The basic annotation of chapters, paragraphs, footnotes, parts in foreign language, parts in italics, or otherwise highlighted parts was introduced following XML/TEI specification accepted for the whole ELTeC collection, the level-1 annotation (https://distantreading.github.io/Schema/eltec-1.html). At level-1, besides some basic TEI structural element, namely <front>, <body>, <back>, , <head>, <p>, <milestone>, and <pb>, some textual elements were introduced as well: <hi>, <foreign>, and <title>.

All these tags were manually introduced during text reading and correction, except for <p>, which was introduced automatically.

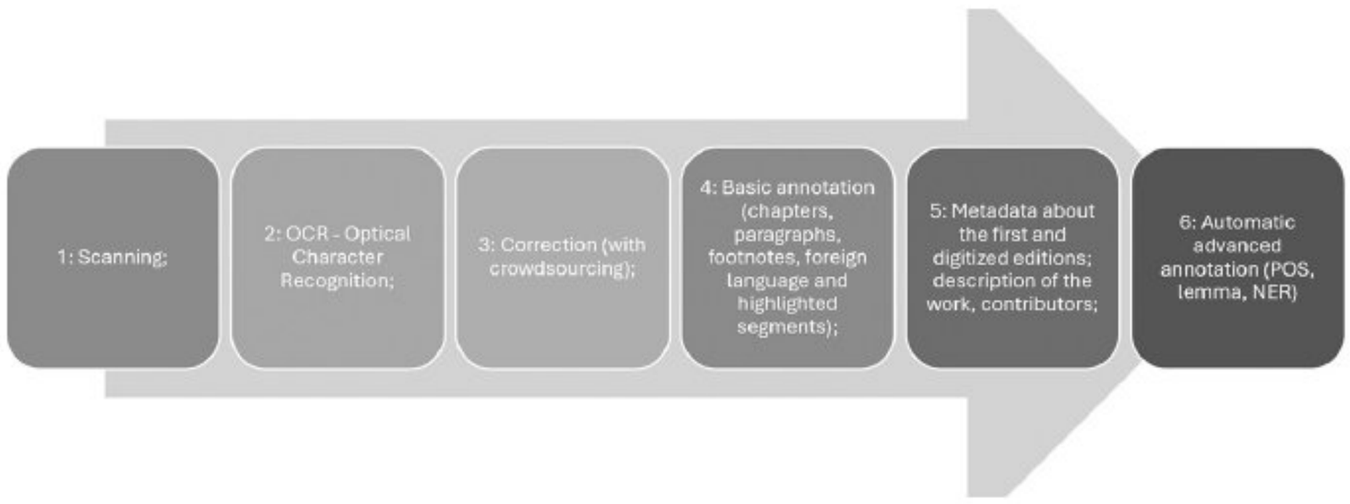


Figure 1: The digitization pipeline for the novels included in SrpELTeC.

Additionally, each novel was equipped with a TEI header with the following obligatory XML elements: <fileDesc> including <titleStmt>, <extent> (size in the number of pages and words), <publicationStmt> (availability and licensing), and <sourceDesc> (source[s] from which it was derived, with the obligatory information about the first edition whether it was used or not for SrpELTeC), <profileDesc> including <langUsage> (language[s] the text was written in), and <textDesc> (text characteristics that serve to check balance criteria of the whole sub-collection). The element <revisionDesc> was added to record all changes made to the file.

Having consistent sub-collection headers enabled the production of various statistics, such as the gender and age of each author (in the time of the publication) as well as statistics related to publishers, publication places, and so on (Krstev, "Serbian Part").

## Automatic annotation of the SrpELTeC Level-2 collection

The ELTeC collection is multi-layered, as the level-2 is built upon level-1 and contains sentence segmentation elements <s>, token elements <w> for words, and <pc> for punctuation. The mandatory attributes for a word token are part-of-speech (@pos), lemma (@lemma), and information about space after a token (@join); since space is not considered a token, its presence or absence after a token is indicated with this attribute. Added to the general XML attribute for the unique identification (@xml:id) is the optional attribute for the more detailed morphosyntactic description (@msd).

The annotation pipeline (Stanković et al., "Annotation") built upon various language resources and tools for Serbian was designed and developed within the JeRTeh Language Resources and Technologies Society. Sentence boundaries were recognized, and sentences were accordingly delimited between <s> and </s> tags using the Serbian specific Unitex transducer (Krstev, *Processing*). The next step was named entity recognition, for which the rule- and lexicon-based SrpNER system was used to enable us to recognize different classes of NEs, such as dates, time, monetary and measurement expressions, geopolitical and personal names, events, and organizations (Krstev et al., "System"). The level-2 tagset contained the following seven classes: PERS, ROLE, LOC, ORG, DEMO, EVENT, and WORK. To map SrpNER using more detailed tags to the corresponding ones from this tagset (Šandrih et al.), various existing and newly developed NER-related tools were integrated into the NER&Beyond online platform (Šandrih Todorović et al.).

TXM tool (Heiden) was used for lemmatization and POS-tagging, adding new information to each token while keeping the existing XML structure intact. The parameter file for TreeTagger (Schmid) was used for the part-of-speech tagging and lemmatization within TXM. The TreeTagger model for the Serbian SrpKor4TaggingTreeTagger (https://live.european-language-grid.eu/catalogue/ld/9296) using the Universal Dependencies tagset (https://universaldependencies.org/u/pos/) was trained on a dataset created from several annotated Serbian texts, SrpKor4Tagging (https://live.european-language-grid.eu/catalogue/corpus/9295). TreeTagger also requires a lexicon and a list of open classes for the training procedure. For this purpose, Serbian morphological dictionaries SrpMD (Krstev, *Processing*) were used to produce the lexicon SrpMD4Tagging in the required format. More about the pipeline can be found in Stanković et al. ("Annotation").

## Some statistics about SrpELTeC

The 100 novels of the core SrpELTeC collection were written by 66 authors: 62 male authors wrote 92 novels, while four female authors (Isidora Sekulić, Jelena Dimitrijević, Draga Gavrilović, and Milica Janković) wrote eight novels. One author (Jaša Ignjatovic) is represented in SrpELTeC with five novels, 12 authors are represented with three novels, 6 authors with two novels, and 47 authors with one novel.

Short novels prevail in SrpELTeC, as 55 texts in the collection have 10,000–50,000 words, followed by 39 medium-length novels

(50.000–100.000 words), while the collection contains only six long novels. The Serbian sub-collection contains 38 canonical novels with a high number of reprints, while the remaining 62 novels were published few times, many of them only once. The first period, 1840–1859, is represented by only two novels, and 18 novels were first published in the years 1860–1879, while the last two periods, 1880–1899 and 1900–1920, are both represented by 40 novels.

The core SrpELTEC contains 4,931,503 words distributed in 368,156 sentences and 149,522 paragraphs across 20,851 pages and 2,329 chapters. The collection also contains 518 quoted segments, 2,873 verses, 853 footnotes, 840 cited works, and 754 phrases in foreign languages.

The average number of words per paragraph is 40, while the average number of words per sentence is 14. The novel with the longest average sentence length, namely 26 words, is *Zločin jedne svekrve* (The Crime of a Mother in Law), while the shortest sentences were used in the novel *Hajduk Stanko* (Haiduk Stanko), the average length being seven words (Stanković et al., "Distant").

A total of 250,340 named entities were tagged in SrpELTeC. Personal names are the most frequent category (PERS 124,338), followed by the persons' professions, positions, or titles (ROLE 82,483), and geopolitical and other urban names (LOC 21,318). The frequencies of other categories are as follows: names of organizations (ORG 1,539), names of inhabitants and ethnic groups, including adjectives derived from geopolitical names (DEMO 19,409), events (EVENT 769), and titles of artistic or professional works (WORK 484). The most frequent first names for men are *Miloš*, *Bože*, *Milan*, *Radiša*, *Stojan*, *Micko*, and *Pera*, while the most frequent names for women are *Jelica*, *Mara*, *Ljubica*, *Darinka*, *Ana*, and *Danica*. The most frequent geopolitical names in SrpELTeC are *Srbija* 'Serbia,' *Beograd* 'Belgrade,' and *Kosovo*. Frequently mentioned countries include *Bosna* 'Bosnia,' *Rusija* 'Russia,' *Turska* 'Türkiye,' *Austrija* 'Austria'; the most frequent inhabited places are *Beograd*, *Carigrad* 'Istanbul,' *Golubac*, *Beč* 'Vienna,' *Niš*, *Skoplje* 'Skopje,' while the most frequent rivers are *Dunav* 'Danube,' *Morava*, *Sava*, and *Drina*.

## Some examples of SrpELTeC usage

In this section we will give some illustrative examples of the use of SrpELTeC in the various domains of research. We will start by demonstrating how comprehensive electronic dictionaries of the Serbian

language can be used to help analyze the content of SrpELTeC. This will be followed by a presentation of a textometric analysis, and finally we will show some different ways of implementing the concept of linked data with the SrpELTeC corpus.

The SrpELTeC corpus was processed using the Unitex/GramLab Multilingual Corpus Processing Suite with the help of a system of electronic morphological dictionaries for the Serbian language (Krstev, "Processing"). One of the research topics was related to eating habits and the language of food. The examples from the corpus show not only what kinds of food were used in the narratives included in SrpELTeC, but also the attitude of the narrated local population towards food, as well as elements of their taste in food as part of their collective identity (Vitas).

One topic of research dealt with the use of alcoholic drinks. Contrary to expectations, *vino* 'wine' was more frequently mentioned than *rakija* 'rakia,' namely 1,181 versus 637 occurrences. If various varieties of wine and rakia are taken into consideration, the ratio remains similar: 1,263 versus 820. Also, whereas less than ten specific brands of *rakia* were mentioned, such as *šljivovica* 'slivowitz,' wine came in more than 30 varieties, including, for instance, *malvasija* 'malvasia.' This type of queries was enabled by semantic markers assigned to entries in the SrpMD e-dictionaries, such as +Drink for drinks. The varieties of wines were categorized in broad categories: *belo vino* 'white wine' (12), *crno vino* 'red wine' (21), and *ružica* or *crveno vino* 'rosé' (2). This shows that red wine was traditionally named *crno vino* 'lit. black wine,' contrary to the current tendency to rename it into *crveno vino* under the influence of English and French. It was also interesting to note that champagne was not unknown to (some) Serbians at that time: it was mentioned 28 times under different names.

Another topic of interest was the evidence of literary works read at the time of SrpELTeC. This information could be retrieved due to the information about the domain assigned to e-dictionaries' personal name entries (in this case, the marker <Dom=Lit> in e-dictionaries). The tag <title> annotating the cited works was also used. Figure 2 presents a few concordance lines: the first column gives the author's name, the novel's title, and publication year, the second column gives the retrieved concordance segment (in Serbian), while its automatic English translation is given in the third column.

| | | |
|---|---|---|
| Šišković, Dragomir: *Jedan od mnogih*, 1920 | Čitala je bez izbora i razbora, sve ruske pisce koji su joj došli do ruku. **Tolstoj, Gogolj, Dostojevski, Gorki, Arcibašev [Sanjin]** redali su se pred njenim očima. Nije prosto znala šta pre da čita. | She read all the Russian writers she could get her hands on without choice or judgment. **Tolstoy, Gogol, Dostoevsky, Gorky, Artzybashev [Sanjin]** lined up before her eyes. She simply did not know what to read first. |
| Jevtić, Stevan J.: *Danica* 1891 | Ti neprestani sanjalo postao si sad najedanput drugi **Demokrit**. Od mene se opet načinio neki novi Epaminonda, „špikovan" **Hajneovom** melanholijom i **Bajronovom** mizantropijom | You never stopped dreaming, now you have suddenly become another **Democritus**. A new **Epaminondas** was made of me again, "spiked" with **Heine's** melancholy and **Byron**'s misanthropy. |
| Komarčić, Lazar: *Prosioci,* 1905 | Znao je „Branka" napamet , i iz „Gorskog Venca" mogao je čitave strane odeklamovati . On vam je mogao, s kraja na kraj , ispričati „Sirotu Bosiljku", „Alpisku Pastirku", „Kasiju Caricu"; „Ezopove Basne" i Dositejeva naravoučenija, tako isto. On je u narodu važio kao neka retka pojava od bistrine, upravo kao neki — seljak filosof. | He knew "**Branka**" by heart, and he could recite whole pages from "**The Mountain Wreath**". He could tell you, from end to end, "**Poor Bosiljka**", "**Alpine Shepherdess**", "**Cassia the Empress**"; "**Aesop's Fables**" and **Dositej's moral lessons**, as well. He was regarded among the people as a rare phenomenon of clarity, precisely as a peasant philosopher. |
| Dimitrijević, Jelena: *Nove*, 1912 | One tri nove nađoše u Fatminoj biblioteci mnoge svoje mile poznanike, pored **Šatobriana** i **Lamartina — Gij de Mopasana**, **Marsela Prevoa**, **Pola Buržea**, i druge. Nađoše i **Flobera**, i **Lotija**... | The three newcomers found many of their dear acquaintances in Fatima's library, in addition to **Chateaubriand** and **Lamartine—Guy de Maupassant**, **Marcel Prévost, Paul Bourget**, and others. They also found **Flaubert** and **Loti**... |

Figure 2: Examples of literary works and their authors mentioned in the SrpELTeC corpus.

Figure 3 lists the most frequently mentioned authors in SrpELTeC, ordered by the frequency: the author (or the title of the work in case of anonymous authors) is given in the first column, the novel that mentions the author is given in the second column, the third column gives an example sentence from that novel, while the fourth column gives the automatic English translation of the sentence. One should note that Goethe, as the author of *Werther*, was by far the most frequently mentioned author, due to the fact that SrpELTeC includes a kind of parody of *Werther*, namely the novel (or long story) *Verter (Werther)* by Laza Lazarević.

| | | | |
|---|---|---|---|
| **Goethe (and his *Werther*)** | Lazarević, Laza: *Verter*, 1881 | Pravi proizvod nemačke poezije. **Gete** video **Rusovljevu** „Novu Eloizu", pa i on napisao „Vertera". | A true product of German poetry. **Goethe** saw Rousseau's "**The New Heloise**", so he also wrote "**Werther**". |
| **Dositej (Obradović)** | Gavrilović, Draga: *Babadevojka*, 1887 | Još mi je bilo slobodno čitati neke prevode od **Fenelona** i **drugih stranih pisaca**, kao i **bibliju** i zlatoustove besede, od **Dositeja Obradovića**, koje su bile još u rukopisu . | I was still free to read some translations by **Fénelon** and other foreign writers, as well as the **Bible** and Zlatumustove besedes, by **Dositej Obradovic**, which were still in manuscript. |
| ***The Bible*** | Kosta Barunčić: *Pastir kralj ili Oslobođenje Srbije*, 1879 | može pesnik naći u dičnoj Srbiji verne obrasce običaja iz **Biblije** i **Omira**. | can a poet find in beautiful Serbia true patterns of customs from the **Bible** and **Homer**. |
| **Tolstoy** | Janković, Milica: *Pre sreće*, 1918 | Čika kaže da je **Mopasan** možda jači, ali je **Tolstoj** bolji. | Uncle says that **Maupassant** may be stronger, but **Tolstoy** is better. |
| **Plutarch** | Popović-Šapčanin, Milorad: *Sanjalo*, 1888 | Učitelj Maksa ode u sobu da ostavi svoga **Plutarha** u školski orman. | Max's teacher went to the room to leave his **Plutarch** in the school closet. |
| **Heine** | Matijašević, Stevan: *Grofica Agneša Janković*, 1897 | — Čitali ste **Hajnea**,... ja proklinjem čas kada sam u ruke uzeo njega i ***Vertera***. | — You read **Heine**, ... I curse the hour when I took him and **Werther** in my hands. |
| **Hugo** | Ljubiša Branković: *Pred Zoru*, 1878 | Razgovarahu se o ***Higovim „Jadnicima"***. | They talked about Hugo's **"Les Misérables"**. |
| | Stevan Sremac: *Pop Ćira i pop Spira*, 1894 | — Ако желите **Работнике на мору** или | — If you want **Workers at Sea** or |
| **Dumas** | Jakov Ignjatović | — Čitala sam od ***Dima „Draj Musketire***" | — I read "**Die drei Musketiere**" by **Dumas** |
| | Svetolik Ranković: *Seoska* učiteljica | — Јеси читала **Монте-Христо**? | — Have you read **Monte Cristo**? |

Figure 3: The most frequently mentioned authors in SrpELTeC.

## Textometry

Besides adding different morphosyntactic annotations, the TXM tool (Heiden) can be used to calculate various text statistics (Krstev et al., "Analysis"), including textometry as a powerful technique for the analysis of text corpora. TXM provides the following qualitative tools: KeyWord in Context concordances of word patterns based on Corpus Query Language; word pattern frequency lists based on tokens, lemmas, parts-of-speech, or structural annotations, including named entities; and word pattern progression graphics. The quantitative analysis tools are based on R packages: factorial correspondence analysis, cluster analysis, specific word patterns analysis, and collocations analysis.

One of the research questions we tackled by using the TXM tool were typical professions of female characters as evidenced in SrpELTeC and their use over time and by different authors. We were able to do that thanks to existing annotations of professions, positions, and titles (tag <role>), and we also had the possibility to include structural tags in queries when using TXM. We began by analyzing the 100 most frequently used roles in the collection and we noticed that 83 refer to men, while 17 refer to women. Among these most frequent roles of men there are numerous professions, including *pop* 'priest,' *kapetan* 'captain,' *doktor* 'doctor,' *učitelj* 'teacher,' *seljak* 'peasant,' *vojnik* 'soldier,' *kaluđer* 'monk,' *sluga* 'servant,' *majstor* 'artisan,' and *pisar* 'registrar,' while roles ascribed to women include only two professions, *sluškinja* 'servent maid' and *učiteljica* '(woman) teacher.' Next, we selected all professions ascribed to women and grouped them in TXM in 7 broad categories, as presented in Figure 4. The figure shows that professions that belong to the category of (manual) workers are the most frequent ones, followed by teachers, where (nursery) governesses are mentioned besides elementary school teachers.

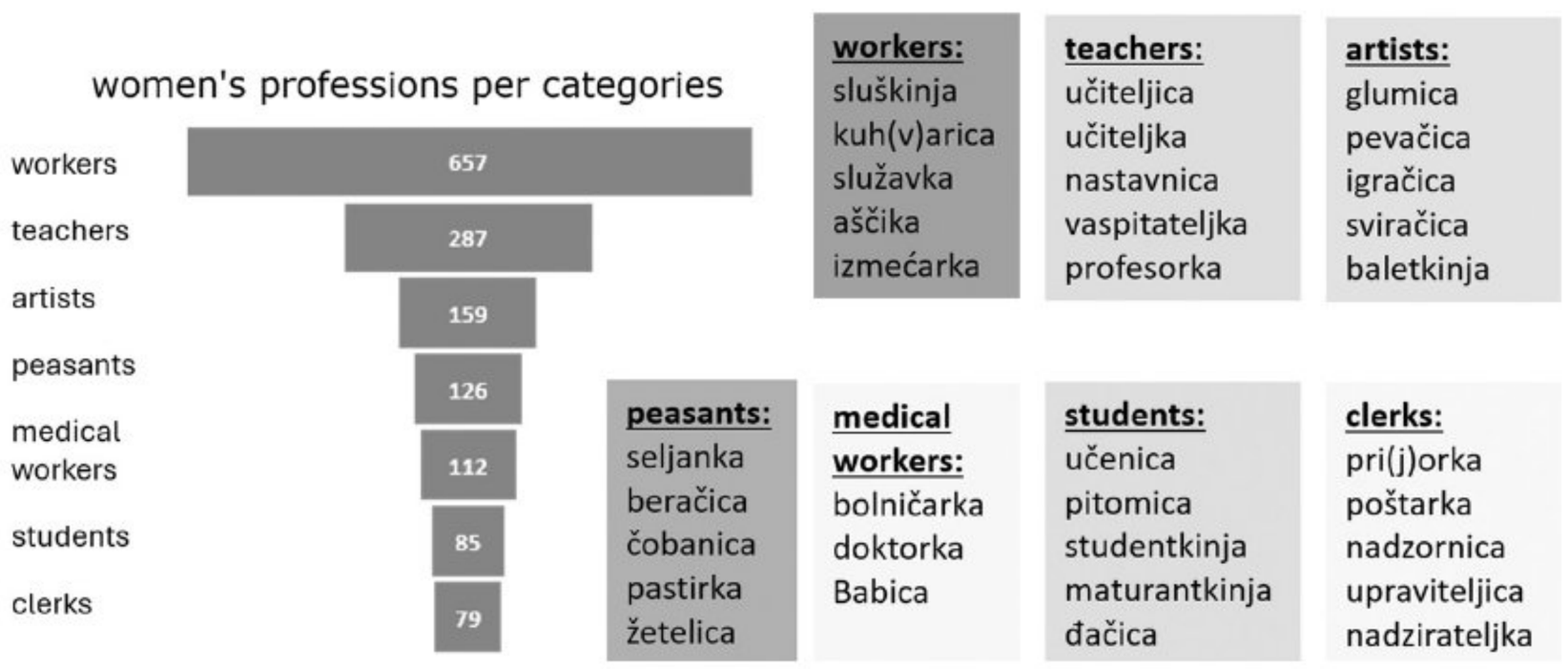


Figure 4: Professions of female characters per categories in SrpELTeC.

Implemented in TXM, progression graphs enabled us to gain insight into the changes in professions of female characters during the period covered by SrpELTeC (see Figure 5). One can observe that workers' professions for female characters were used more or less uniformly throughout the collection. The same holds for peasants, but less frequently. On the other hand, women teachers appeared later (there is a steep rise after one million corpus words), which is due to three novels written by Draga Gavrilović in which teacher maids are the main characters. After that, teacher maids did not occur in novels for a long time (hence the straight line after approximately one million words). There is a slight rise of women artists, clerks, and medical workers by the end of the period; however, in the same period, female students did not seem to be interesting to novelists.

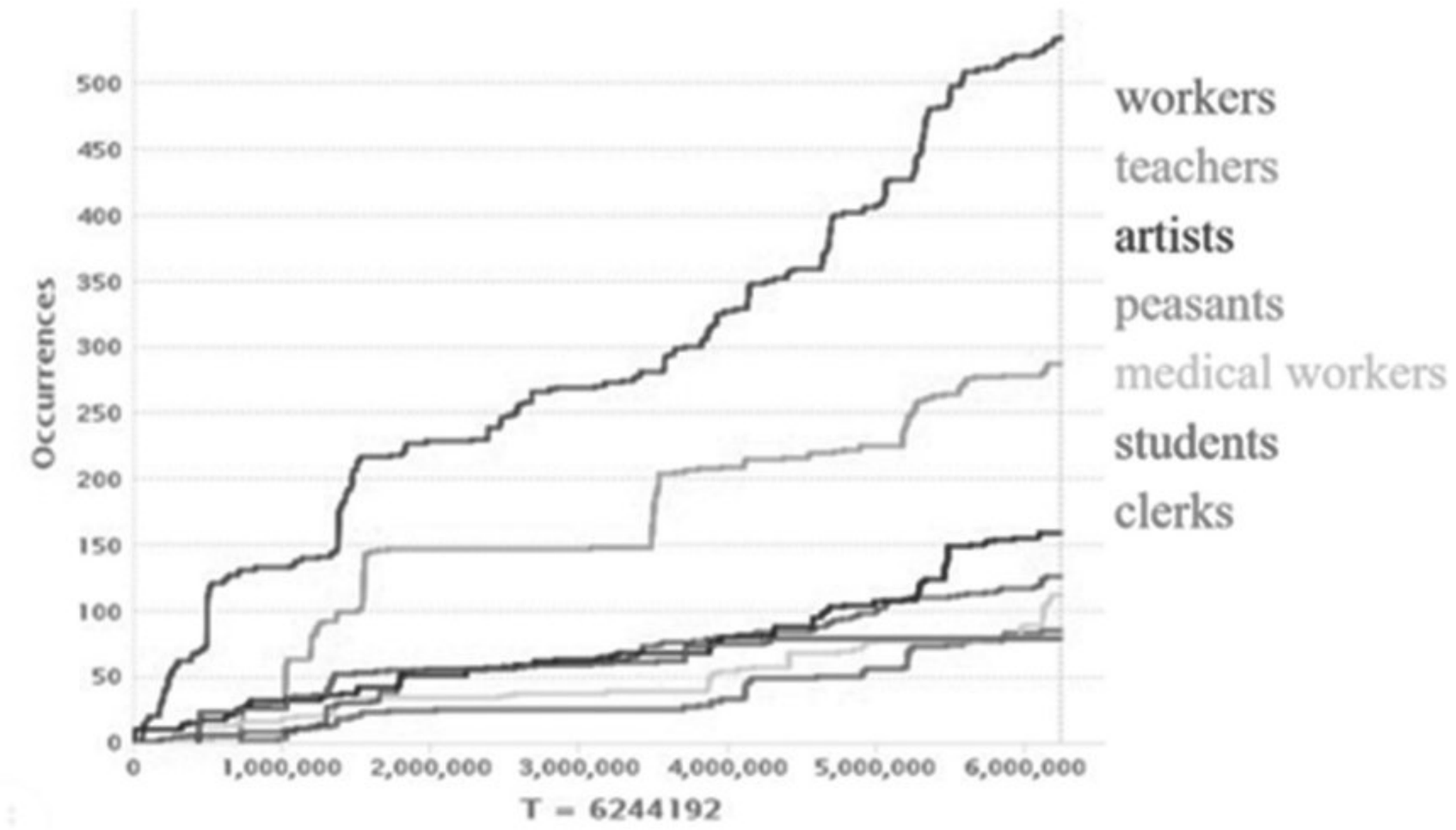


Figure 5: Changes in professions of female characters over time in SrpELTeC; the numbers of corpus words are given on the *x*-axis, but as the novels in SrpELTeC are sorted according to the year of the first publication, this axis also refers to time.

With the textometric approach to the corpus it is possible to recognize some specific entities or characteristics along with their high or low representation in certain parts of the corpus. The specificity score based on the hyper-geometric distribution shows the probability of a lexical unit occurring in a particular part of the corpus. The TXM also provides a graphic representation of the specificity distribution of the selected units. Specificity score values higher (positive) or lower (negative) than expected express a more or less represented lexical unit or pattern

(Heiden). This is illustrated in Figure 6. The corpus was separated in two partitions: works written by women and works written by men. The occurrence of professions of female characters in the two partitions was analyzed, showing that women teachers tended to be written about by female authors, while women peasants were more often thematized by male authors.

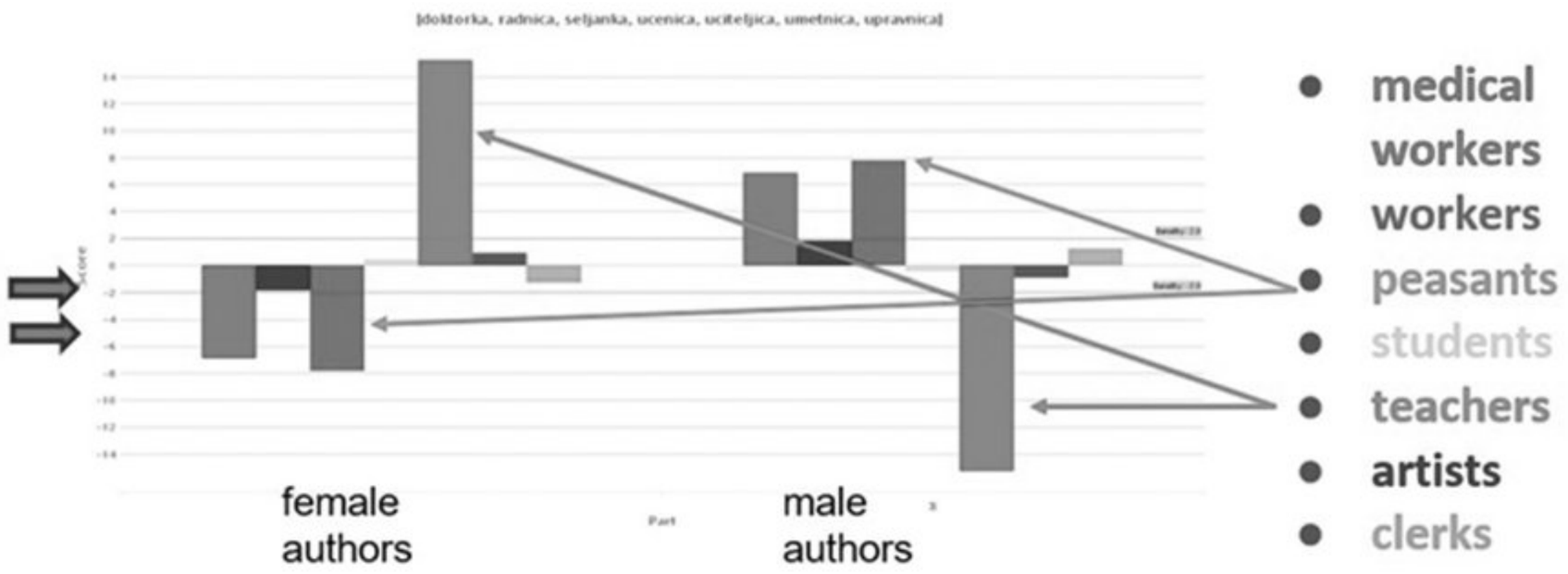


Figure 6: The specificity of the appearance of professions of female characters by the gender of authors in SrpELTeC.

Having in mind that SrpELTeC will be used for various kinds of lexical and linguistic research, the use of diverse tools and methodologies was provided to make it available through various channels, thus hopefully meeting the needs of different types of users. Three platforms on which these novels are published, namely, Udaljeno čitanje, Aurora, and Sketch Engine, are presented in Stanković et al. ("SrpELTeC"). The Udaljeno čitanje platform is intended for readers who would like to see the original print as a picture while reading the digitized version. The Aurora portal provides researchers of Serbian literature and other interested users with a detailed inspection of the novel's vocabulary, enabling text browsing, concordances, and frequency lists. The Sketch Engine (Kilgarriff et al., "Sketch Engine"; Kilgarriff et al., "Sketch Engine: Ten Years On") is a platform for corpora management and exploration, as well as for analyzing texts to identify what is typical in a language and what is a rare, unusual, or emerging usage. A NoSketch Engine node is installed and maintained by JeRTeh, offering access to several monolingual and bilingual corpora. The SrpELTeC corpus can be freely accessed and searched using Corpus Query Language, and registration is possible without any special conditions.

## Linguistic Linked Open Data

With the development of Linguistic Linked Open Data, interest in formalizing the bridge between digital humanities and web-centered linguistic data science has been intensified, although mainly with a focus on lexical data. In the digital humanities community, TEI XML-based standards represent the prototypical publishing approach and have been criticized for not establishing a sufficient degree of interoperability and synchronization with formal semantics and web standards such as RDF and OWL.

Preparing the data on the SrpELTeC novels for Wikidata and linking Wikidata to various applications started as a manual process. The opportunity to speed up this process was seen in using information already encoded in the TEI header of each novel. Data on 700 novels in seven languages from the ELTeC collection were introduced in Wikidata as part of the WikiELTeC project (Ikonić Nešić et al., "From ELTeC"). Since the automation of the process of data preparation and import was envisaged, the different solutions were analyzed, and finally the synergy of OpenRefine and QuickStatements tools was chosen as the best choice (Ikonić Nešić et al., "Serbian ELTeC"). WikiELTeC was semi-automatically populated from <TeiHeader> using OpenRefine, QuickStatements, and custom-made procedures; after the extraction of metadata from headers, the mapping with Wikidata schema was defined in OpenRefine and predicates (properties) that connected subjects and objects in RDF triples were specified in the OpenRefine table header. Each statement for a subject has a property and a value that can become a Wikidata item, an external URL, or a literal (string). After consolidation in OpenRefine, RDF triplets were imported in Wikidata using QuickStatements.

Several groups of data were added or improved, including authors, publishers, metadata about novels, the novels' printed and electronic editions, including SrpELTeC, main characters and their relations, and novels' settings. As a result, 71 authors and 120 novels from the core and the extended SrpELTeC collections were represented in Wikidata, comprising, together with associated items for first editions and digital SrpELTeC editions, approximately 3,500 statements. Metadata about the novels were automatically imported, while main characters and their relations were added manually by volunteers, mostly students at the University of Belgrade and members of JeRTeH.

Each novel's metadata item is linked with an appropriate metadata instance for electronic edition (Q59466853), first edition (Q10898227), print edition (Q59466300), and digital edition (Q1224889), using the

property (P747: has edition or translation), and every item of an edition must be connected with a corresponding item for a novel with the inverse property (P629: edition or translation of). The list of all properties is documented in WikiProject_ELTeC.

A more detailed description of SrpELTeC is provided by introducing main characters and the places mentioned in the novels. This was done within the Serbian WikiProject WikiELTeC (https://sr.wikipedia.org/wiki/Википедија:Википројекат_WikiELTeC), in the scope of which numerous SPARQL queries were provided for the retrieval, analysis, and diverse visualization options of the SrpELTeC data. Main characters of the novels were linked not only with their family and social relations, but also with actors who interpreted them in films and TV series. The understanding of the corpus got a new dimension with provided wikidata supported by SPARQL queries. For example, browsing novels written by Borisav Stanković (Q370392), one can start with the following query and continue by browsing the graph as presented in Figure 7:

```
#defaultView:Graph
SELECT DISTINCT ?author ?authorLabel ?authorImage ?novel ?novelLabel
?image
WHERE {
VALUES ?author { wd:Q370392}
?novel wdt:P50 ?author;
wdt:P18 ?image;
wdt:P747 ?edition.
?edition wdt:P1433 ?collection.
OPTIONAL {?author wdt:P18 ?authorImage}
SERVICE wikibase:label
{ bd:serviceParam wikibase:language «sr,[AUTO_LANGUAGE],en». }}
```

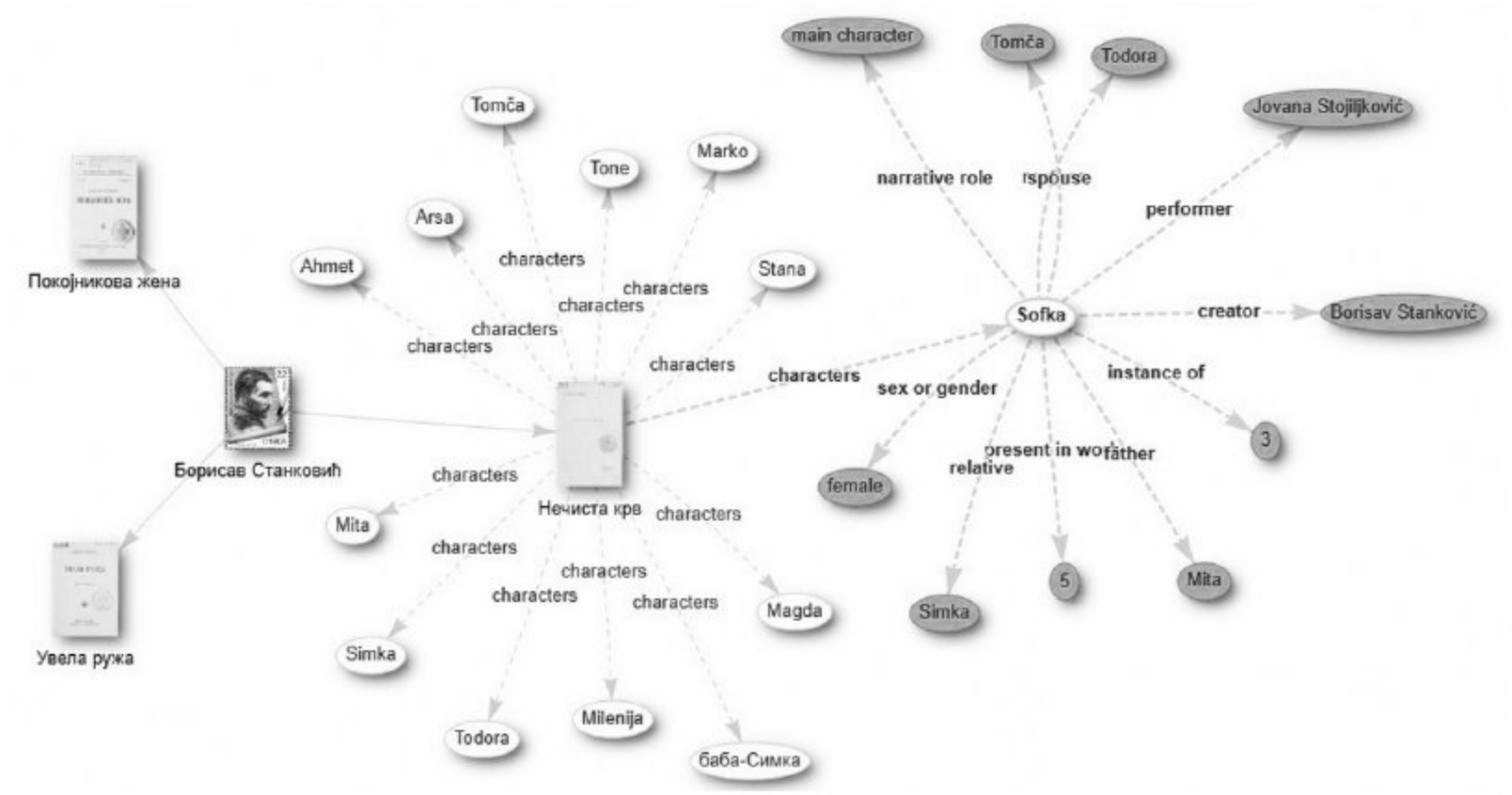


Figure 7: The knowledge graph for one author (Borisav Stanković) obtained by the presented query (at the top), his novels, and one novel's (*Nečista krv* [Impure Blood]) related data.

Introducing characters and places mentioned in novels in WikiData enables linking their occurrences in the texts of SrpELTeC. The INCEPTION tool was used for this, where one can start with the manual annotation while the system learns to make predictions as annotation advances. An example is presented in Figure 8: rivers *Ništava* (Q583062) and *Dunav* (Q1653) are annotated as locations (LOC) and linked to their corresponding Wikidata QIDs. The same figure demonstrates that professions and titles, annotated with the ROLE label, are also linked to WikiData, namely *trgovac* 'merchant' (Q215536) and *pisac* 'writer' (Q36180). Current activities are focused on training a model for automatic linking of recognized named entities.

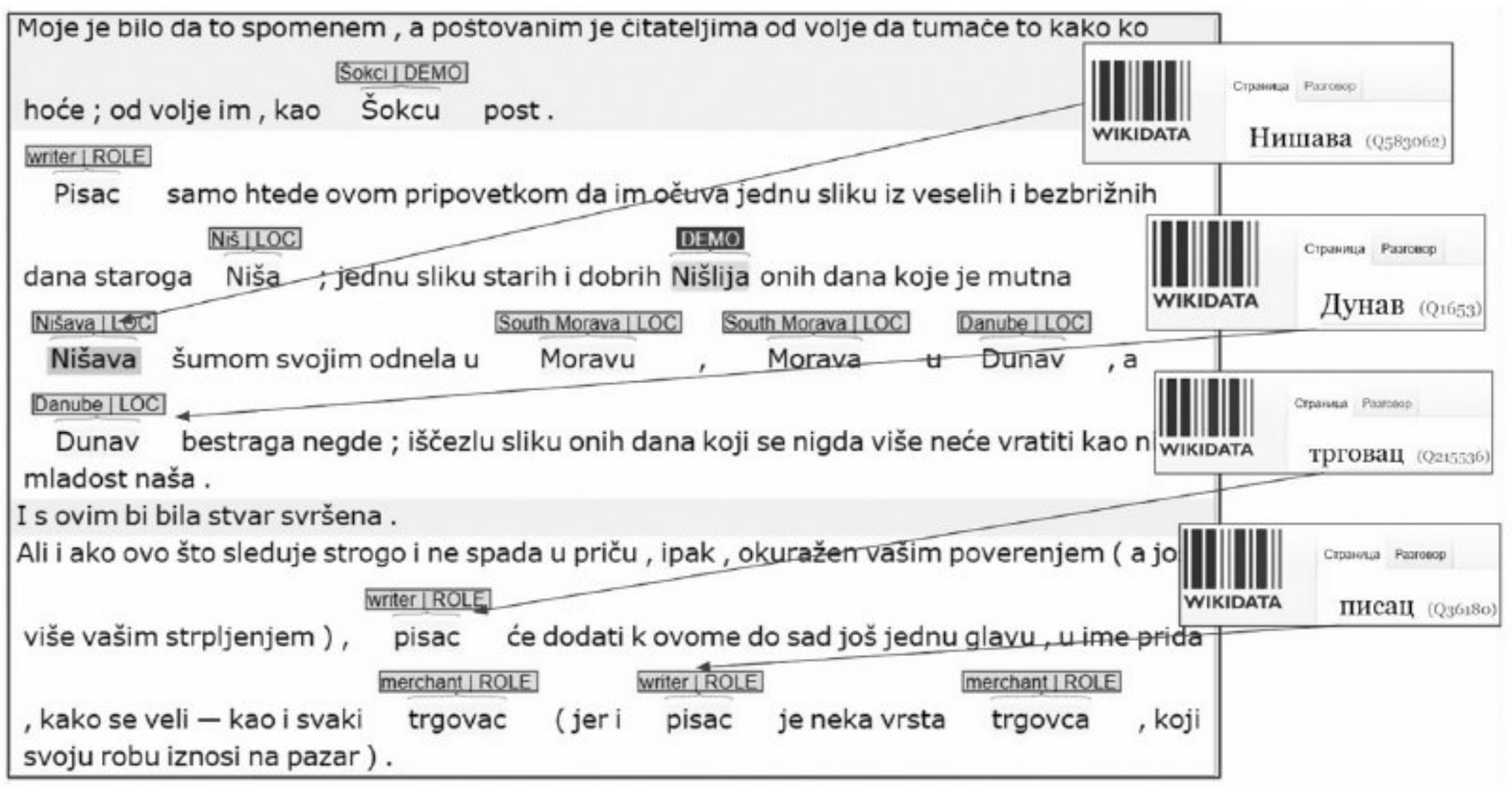


Figure 8: A text in the INCEPTION platform with examples of entity linking with Wikidata.

The linked data version of SrpELTeC using NLP Interchange Format (NIF) was produced within the COST Action NexusLinguarum–European Network for Web-centred Linguistic Data Science (CA18209). NIF is designed to facilitate the integration of NLP tools in knowledge extraction pipelines; it provides support for part-of-speech tagging, lemmatization, and entity annotation, enabling ELTeC level-2 layer transformation (Stanković et al., "Towards"). Several ontologies were consulted to use equivalents of named entity types: OLIA, DBpedia, Wikidata. For Wikidata, the following mapping was specified: wd:Q5 (PERS), wd:Q7884789 (LOC), wd:Q43229 (ORG), wd:Q1656682 (EVENT), wd:Q28640 (ROLE), wd:Q217438 (DEMO), and wd:Q386724 (WORK).

We should note that recognized named entities have not been linked with Wikidata or DBpedia items yet, as they are only marked and classified into one of seven predefined types. The Apache Jena Fuseki server was used for testing the Serbian ELTeC NIF corpus SPARQL at the JeRTeh site (http://fuseki.jerteh.rs/#/dataset/SrpELTeC/query). A SPARQL query presented in Figure 9 illustrates the retrieval of the most frequently used nouns in novels written by Jakov Ignjatović (wd:Q570913), that is, *kuća* 'house' (275), *otac* 'father' (208), *dan* 'day' (144), *mati* 'mother' (140), *godina* 'year' (127), and *ruka* 'hand' 123.

SPARQL Endpoint /SrpELTeC/sparql Content Type (SELECT)

```
PREFIX nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
PREFIX xsd: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT ?lemma (COUNT(?lemma) AS ?count)
   WHERE {
    ?subject nif:lemma ?lemma ;
       nif:posTag "NOUN"^^<http://www.w3.org/2001/XMLSchema#string>;
       nif:referenceContext ?novelid.
    # Jakov Ignjatović
    ?novelid dc:creator wd:Q570913.
   }
GROUP BY   ?lemma
ORDER BY desc(?count)
```

Table Response 2514 results in 3.109 seconds

| | lemma | count |
|---|---|---|
| 1 | кућа | "275"^^<http://www.w3.org/2001/XMLSchema#integer> |
| 2 | отац | "208"^^<http://www.w3.org/2001/XMLSchema#integer> |
| 3 | дан | "144"^^<http://www.w3.org/2001/XMLSchema#integer> |
| 4 | мати | "140"^^<http://www.w3.org/2001/XMLSchema#integer> |
| 5 | година | "127"^^<http://www.w3.org/2001/XMLSchema#integer> |
| 6 | рука | "123"^^<http://www.w3.org/2001/XMLSchema#integer> |

Figure 9: The Fuseki node with the SrpELTeC edition as linked data using NIF: an example of a query.

## Conclusion

It is our belief that the developed platforms and different formats and editions of digital versions of novels will contribute to raising the visibility of SrpELTeC as a valuable Serbian language resource for linguists and literary scholars. Additionally, they will shed light on a relatively unknown period of Serbian literary history. The SrpELTeC corpus can help paint a portrait of life in Serbia in the second half of the nineteenth century and early twentieth century. Besides investigating eating habits, literature read, and professions practiced, more complex issues, such as the position of women in the Serbian society, education of children, cultural habits, medical treatments, means of travel, or attitudes towards so-called others, could be explored, opening possibilities for further analysis of the corpus. To this end, multi-word expressions and other additional layers will be added to the collection. Finally, valuable data should be added to the header information by literary scholars, including the dialects and pronunciations used in novels, the backgrounds of the authors (e.g., whether an author's birthplace was in the Turkish Empire or Austro-Hungarian Empire), and the generic features of the novels.

WORKS CITED

Burnard, Lou, et al. "In Search of Comity: TEI for Distant Reading." *Journal of the Text Encoding Initiative*, vol. 14, 2021, pp. 1–9, https://doi.org/10.4000/jtei.3500. Accessed 24 Jan. 2024.

Deretić, Jovan. *Istorija srpske književnosti*. Belgrade, Nolit, 1983.

Deretić, Jovan. *Srpski roman 1800–1950*. Belgrade, Nolit, 1981.

Heiden, Serge. "The TXM Platform: Building Open-source Textual Analysis Software Compatible with the TEI Encoding Scheme." *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. Vol. 2. No. 3*, edited by Ryo Otoguro et al., Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010, pp. 389–398, https://aclanthology.org/Y10-1044. Accessed 24 Jan. 2024.

Ikonić Nešić, Milica, et al. "From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)." *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, edited by Thierry Declerck et al., European Language Resources Association, Paris, 2022, pp. 7–16, https://aclanthology.org/2022.ldl-1.2/. Accessed 24 Jan. 2024.

Ikonić Nešić, Milica, et al. "Serbian ELTeC Sub-Collection in Wikidata." *Infotheca*, vol. 21, no. 2, 2021, pp. 60–87, https://doi.org/10.18485/infotheca.2021.21.2.4. Accessed 24 Jan. 2024.

Kilgarriff, Adam, et al. "The Sketch Engine." *Proceedings of the Eleventh EURALEX International Congress*, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, 2004, pp. 105–115, https://euralex.org/publications/the-sketch-engine/. Accessed 24 Jan. 2024.

Kilgarriff, Adam, et al. "The Sketch Engine: Ten Years On." *Lexicography*, vol. 1, no. 1, 2014, pp. 7–36, https://doi.org/10.1007/s40607-014-0009-9. Accessed 24 Jan. 2024.

Krstev, Cvetana. *Processing of Serbian: Automata, Texts and Electronic Dictionaries*. Belgrade, Faculty of Philology of the University, 2008.

Krstev, Cvetana. "The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata." *Infotheca*, vol. 21, no. 2, 2021, pp. 26–42, https://doi.org/10.18485/infotheca.2021.21.2.2. Accessed 24 Jan. 2024.

Krstev, Cvetana, et al. "A System for Named Entity Recognition Based on Local Grammars." *Journal of Logic and Computation*, vol. 24, no. 2, 2014, pp. 473–489.

Krstev, Cvetana, et al. "Analysis of the First Serbian Literature Corpus of the Late 19th and Early 20th Century with the TXM Platform." *DH_BUDAPEST_2019*, Eötvös Loránd University, Centre for Digital Humanities, 2019, pp. 36–37.

Milisavac, Živan, editor. *Pripovedači*. Novi Sad / Belgrade, Matica srpska / Srbska knjizhevna zadruga, 1972.

Moretti, Franco. "Conjectures on World Literature." *New Left Review*, vol. 1, 2000, pp. 54–68.

Odebrecht, Carolin, et al. *European Literary Text Collection (ELTeC): April 2021 Release with 14 Collections of at Least 50 Novels (v1.1.0)*. Zenodo, 2021, https://doi.org/10.5281/zenodo.4662444. Accessed 24 Jan. 2024.

Patras, Roxana, et al. "Thresholds to the 'Great Unread': Titling Practices in Eleven ELTeC Collections." *Interférences Littéraires / Literaire Interferenties*, vol. 25, 2021, pp. 163–187, http://interferenceslitteraires.be/index.php/illi/article/view/1102. Accessed 24 January 2025.

Schmid, Helmut. "Improvements in Part-of-speech Tagging with an Application to German." *Natural Language Processing Using Very Large Corpora*, edited by Susan Armstrong et al., Springer Netherlands, 1999, pp. 13–25.

Schöch, Christof, et al. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives." *Modern Languages Open*, vol. 1, 2021, https://doi.org/10.3828/mlo.v0i0.364. Accessed 24 Jan. 2024.

Stanković, Ranka, et al. "Annotation of the Serbian ELTeC Collection." *Infotheca*, vol. 21, no. 2, 2021, pp. 43–59, https://doi.org/10.18485/infotheca.2021.21.2.3. Accessed 24 Jan. 2024.

Stanković, Ranka, et al. "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection." *LREC 2022 Conference Proceedings*, edited by Nicoletta Calzolari et al., European Language Resources Association, Paris, 2022, pp. 3337–3345, https://aclanthology.org/2022.lrec-1.356/. Accessed 24 Jan. 2024.

Stanković, Ranka, et al. "SrpELTeC on Platforms: Udaljeno čitanje, Aurora, noSketch." *Infotheca*, vol. 21, no. 2, 2021, pp. 136–153, https://doi.org/10.18485/infotheca.2021.21.2.7. Accessed 24 Jan. 2024.

Stanković, Ranka, et al. "Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data." *Language, Data and Knowledge 2023*, edited by Sara Carvalho et al., NOVA CLUNL, Lisbon, 2023, pp. 180–191.

Šandrih, Branislava, et al. "Development and Evaluation of Three Named Entity Recognition Systems for Serbian–The Case of Personal Names." *RANLP 2019: Natural Language Processing in a Deep Learning World: Proceedings*, edited by Galia Angelova et al., INCOMA Ltd., Shumen, 2019, pp. 1060–1068, https://aclanthology.org/R19-1122/. Accessed 24 Jan. 2024.

Šandrih Todorović, Branislava, et al. "Serbian NER&Beyond: The Archaic and the Modern Intertwinned." *RANLP 2021: Deep Learning for Natural Language Processing Methods and Applications*, edited by Galia Angelova et al., INCOMA Ltd., Shumen, 2021, pp. 1252–1260, https://aclanthology.org/2021.ranlp-1.141/. Accessed 24 Jan. 2024.

Trtovac, Aleksandra, et al. "The Serbian Part of the ELTeC–From the Empty List to the 100 Novels Collection." *Infotheca*, vol. 21, no. 2, 2021, pp. 7–25, https://doi.org/10.18485/infotheca.2021.21.2.1. Accessed 24 Jan. 2024.

Vitas, Duško. "From Onions to Champagne–Food and Drink in the SrpELTeC Corpus." *Infotheca*, vol. 21, no. 2, 2021, pp. 88–118, https://doi.org/10.18485/infotheca.2021.21.2.5. Accessed 24 Jan. 2024.

# SrpELTeC: srbski literarni korpus za oddaljeno branje

Ključne besede: digitalna humanistika / srbska književnost / besedilni korpusi / oddaljeno branje / povezani podatki / prepoznavanje imenskih entitet / besedilna analitika

V članku je predstavljen korpus SrpELTeC, ki je nastal v okviru COST akcije Distant Reading for European Literary History (CA16204). Vsi romani v SrpELTeC-u so bili izbrani, pripravljeni in označeni po skupnih načelih, ki veljajo za vse jezikovne zbirke v European Literary Text Collection (ELTeC). Opisani so izzivi in rešitve pri pripravi SrpELTeC. Vsi romani so bili ročno kodirani skladno s parametri XML-TEI in opremljeni z bogatimi metapodatki ter strukturnimi opombami. Avtomatično označevanje je vključevalo oblikoskladenjske oznake, lematizacijo in imenske entitete, ki so temeljile na virih za obdelavo naravnega jezika, ki jih je razvilo in jih vzdržuje Društvo za jezikovne vire in tehnologije JeRTeh. Integracija SrpELTeC z Wikidata je bila podprta z nizom poizvedb SPARQL za pridobivanje metapodatkov z različnimi možnostmi vizualizacije. V okviru nedavnih dejavnosti v okviru COST akcije NexusLinguarum – European Network for Web-centred Linguistic Data Science (CA18209) je bila ustvarjena povezana podatkovna različica SrpELTeC z uporabo NLP izmenjevalnega formata. Vse različice SrpELTeC so prosto dostopne pod licenco CC-BY.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.163.41.09:004
DOI: https://doi.org/10.3986/pkn.v47.i2.03