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

EDITOR-IN-CHIEF
Prof. Cvetana Krstev, PhD cvetana@matf.bg.ac.rs
Faculty of Philology, Department for Library and Information Science

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

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

EDITORIAL BOARD:
Prof. Aleksandra Vraneš, PhD, Prof. Aleksandar Jerkov, PhD, Prof. Biljana Dojčinović, PhD, *Faculty of Philology, University of Belgrade*; Prof. Elisabeth Burr, PhD, *Institut für Romanistik, Universität Leipzig*; Prof. Vladan Devedžić, PhD, *Faculty of Organization Sciences, University of Belgrade*; prof. Milena Dobreva, PhD, *Faculty of Media and Knowledge Sciences, University of Malta*; Tomaž Erjavec, PhD, *Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana*; Prof. Svetla Koeva, PhD, *Institute for Bulgarian Language, Bulgarian Academy of Sciences*; Prof. Denis Maurel, PhD, prof. Agata Savary, PhD, *Université Francois Rabelais de Tours*; Prof. Ivan Obradović, PhD, *Faculty of Mining and Geology, University of Belgrade*; Prof. Gordana Pavlović Lažetić, PhD, prof. Duško Vitas, PhD, *Faculty of Mathematics, University of Belgrade*; Prof. Katerina Zdravkova, PhD, *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje*

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

Belgrade, Vol. 19, No. 1, September 2019

WEB PORTAL:
Jelena Andonovski
University Library "Svetozar Marković"

LECTORS FOR ENGLISH:
Tanja Ivanović
Ministry of European Integration

LECTOR FOR SERBIAN:
Svjetlana Đelić
University Library "Svetozar Marković"

DESIGN AND PREPRESS:
Branislava Šandrih
and **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 Infotheca
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infotheca@unilib.rs

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Contents

Creation and Analysis of the Yugoslav Rock Song Lyrics Corpus from 1967 to 2003[1]

UDC 811.163.41'322
DOI 10.18485/infotheca.2019.19.1.1

ABSTRACT: The paper analyses the process of creation and processing of the Yugoslav rock song lyrics corpus from 1967 to 2003, from the theoretical and practical perspective. The data have been obtained and XML-annotated using the Python programming language and the libraries lyricsmaster/yattag. The corpus has been preprocessed and basic statistical data have been generated by the XSL transformation. The diacritic restoration has been carried out in the Slovo Majstor and LeXimir tools (the latter application has also been used for generating the frequency analysis). The extraction of sociocultural topics has been performed using the Unitex software, whereas the prevailing topics have been visualised with the TreeCloud software.

KEYWORDS: corpus linguistics, Yugoslav rock and roll, web scraping, natural language processing, text mining.

PAPER SUBMITTED: 15 April 2019
PAPER ACCEPTED: 19 June 2019

Ljudmila Petković
ljudmila.petkovic@gmail.com
University of Belgrade
Belgrade, Serbia

[1] This paper originates from the author's Master's thesis "Creation and Analysis of the Yugoslav Rock Song Lyrics Corpus from 1945 to 2003", which was defended at the University of Belgrade on March 18, 2019. The thesis was conducted under the supervision of the Prof. Dr Ranka Stanković, who contributed to the topic's formulation, with the remark that the year of 1945 was replaced by the year of 1967 in this paper.

1 Introduction

1.1 The Phenomenon of the Yugoslav rock and roll

Music analysts, sociologists and anthropologists unanimously agree that the Yugoslav rock and roll (also known as the *Yu rock*) had left a deep impact on the territory of the former Yugoslavia. It is a musical style that was conceived in 1961, along with the appearance of the bands Uragani, Bijele Strijele and Siluete, while during the same decade the groups Crveni Koralji (1962), Zlatni Dečaci (1962) and Korni Grupa (1968) were formed (Janjatović, 1998). The Yu rock had been developing in parallel with the then-flourishing British *beat* scene (Раковић, 2011), made famous by the bands The Beatles and The Rolling Stones (Cooper and Cooper, 1993). In the late fifties and early sixties, the rock and roll in Yugoslavia had been equated with the rock and roll/twist dance, and not with the music genre per se (Раковић, 2011).

In 1963, the rock and roll in the Socialist Federal Republic of Yugoslavia (abbr. SFRY) had acquired the status of genre, which since then had been primarily performed with electric guitars, bass and drums (Раковић, 2011). Classic rock and rockabilly had represented very popular musical forms among the Yugoslavs, as evidenced by the covers of Elvis Presley, Buddy Holly, Chuck Berry and the like. (Арсенијевић et al., 2016). The 1970s had brought the influence of the hippie movement, with additional genre layering and the emergence of hard rock, progressive rock, art rock, and similar subgenres. The end of the 1970s and the early 1980s had been marked by the advent of punk rock and new wave, based on the equivalent seminal forms originating from the USA and the UK (Арсенијевић et al., 2016). In fact, the "Yugo-rock" has found its place in a society that had enthusiastically embraced the products of Western culture, both via domestic journals (e.g. *Rock* and *Džuboks*) and foreign radio programmes broadcast by Radio Luxembourg, military and pirate stations, as well as in form of films from the West. Also, Western fashion trends has also been followed, which had dictated, among other things, the men's wearing of long hair or the girls' wearing of mini skirts (Раковић, 2018).

Nevertheless, what made Yu-rock particularly distinctive were the libertarian lyrics, characteristic of punk and new wave music. Namely, the lyrics content was often politically engaged, ironic or vulgar, and therefore treated as unfit for the then-Yugoslav social circumstances (Гајић, 2018). The systematisation of the (self-)censorship cases in Yu-rock represents evidence

that this was not an isolated phenomenon.[2] On the other hand, numerous songs from that era extolled the homeland, while in the others there were allusions to the partisans and the brigade (Божиловић, 2016). Božilovic adds that the core of Yu-rock was constituted by rock musicians' rebellion and fight for democracy, which had been encapsulated into unconventional, simple, direct and improvised musical forms and lyrics. However, Yu-rock, like its Western predecessor, also gave birth to the artists who sublimed their personal life attitudes into lyrics with lyrical, philosophical and introspective tone (e.g. the bands Azra, Idoli or Ekatarina Velika).

1.2 State of the Art

To the best of the author's knowledge, the studies of Yugoslav rock and roll have until now been primarily based on theoretical considerations, without application of computer technologies (some of such works are listed in the subsection 1.1). Zörnig et al. (2016) have conducted the only research that dealt with the quantitative analysis of song lyrics from the Yugoslav rock and roll era. Part of the corpus in the aforementioned work contains lyrics of the band Riblja Čorba and his frontman Bora Đorđević. The same paper presents the method of calculating the word frequencies and lexical variability using the *relative repeat rate* and *h-point* measures in the first place, which allowed the classification of corpus texts with multivariate analysis.

On the other hand, automatic collecting of song lyrics from the web, their electronic processing and analysis are obtaining more and more attention, as evidenced by the vast number of projects and scientific studies. There is a large number of international song lyrics corpora, e.g. The Million Song Dataset[3] (Bertin-Mahieux et al., 2011), which can be used in the computational text analysis researches. So far, the lists of the most frequent words in the annotated corpora of pop (Kreyer and Mukherjee, 2007) and rock song lyrics (Falk, 2013) have been generated. Topic modeling techniques were applied to the librettos of the Beijing opera (Zhang et al., 2017), as well as to the corpus of song lyrics harvested from the website SongMeanings (Lukic, s.d.).

When it comes to song lyrics from the domain of rock or related genres, the computational analysis of their content represents a promising and highly developed scientific practice. Some interesting results that these techniques

[2] See the website Balkanrock.
[3] The Million Song Dataset (on-line).

have produced are the structure extraction of The Beatles' song lyrics (Mahedero et al., 2005) or plotting the song lyrics written by Paul McCartney and John Lennon on the *emotion clock* circumplex (Whissell, 1996). The research carried out by Petrie et al. (2008) is another example of interest in computational processing of The Beatles' song lyrics. The given research is concerned with the change of the prevailing mood in their lyrics, while the use of stylometric analysis reveals stylistic similarities and differences between Lennon, McCartney, and George Harrison, as songwriters.

Falk (2013) has brought the results of a diachronic analysis of linguistic peculiarities in the corpus of international rock song lyrics from 1950 to 1999, based on the most frequent words used in each decade. Haslam (2017) has traced the evolution of thematic motives in the songs of the singer-songwriter Leonard Cohen, which has been visualised in the form of *word clouds*. From the standpoint of corpus linguistics, Taina et al. (2014) has explored the linguistic features in the song lyrics that distinguish heavy metal subgenres from each other. The tools used in psychometric research in order to calculate the textual cohesion (Coh-Metrix)[4] and to discover emotional, cognitive, and structural components in texts (Linguistic Inquiry and Word Count)[5] have been implemented in the comparative analysis of rock, folk, country, punk and grunge song lyrics (Lightman et al., 2007). The aim of the mentioned research was to identify differences in the writing style between artists who committed suicide and those who were not suicidal. Thematically more remote, but methodologically close to the present research is the paper dealing with the techniques for the automatic extraction of multi-word expressions in the lyrics of one ancient religious Hindu poem using local grammars in Unitex[6] software (Stein, 2012).

In all likelihood, the analysis of Yugoslav rock song lyrics from the aspect of computational linguistics seems to be an underdeveloped field of research. For that reason, this paper seeks to examine Yu-rock songs through the prism of interdisciplinarity, in order to reinforce the existing interpretations of the genre with the results produced by computational corpus processing tools.

1.3 Theoretical-Methodological Framework

One of the focal points of the current research is the application of corpus linguistics techniques. Corpus linguistics is a developed scientific methodol-

[4] Coh-Metrix (on-line).
[5] Linguistic Inquiry and Word Count (abbr. LIWC) (on-line).
[6] Unitex/GramLab (on-line).

ogy, while many scholars also consider it a discipline, theory, paradigm and a tool (Taylor, 2008). This methodological approach deals with the handling of structured, machine-readable and purposely chosen texts, which represent the basis for analysing various aspects of language and its usage. In the general context of corpus linguistics, the "purposely chosen texts" mean the collection of certain textual units sampled for the purpose of achieving representativeness in corpus construction. Frequency of a particular word or phrase usage, their retrieval as keywords in context by generating a concordance, or metadata extraction (e.g. name of the author of the text, date of publication, language in which the text is written, etc.) from an annotated corpus are just some of the possible tasks of corpus linguistics (McEnery and Hardie, 2012).

The data for the Yu-rock song lyrics corpus have been collected by using the web scraping[7] method. Algorithms for automatic collection, preprocessing and annotating the corpus in accordance with the XML syntax have been implemented in the Python programming language using the `lyricsmaster`, `xml.sax.utils` and `yattag` libraries. The XSLT transformation of the XML document into XHTML format allows the overview of the corpus statistics.[8] Automatic diacritic restoration has been conducted using the Slovo Majstor[9] and LeXimir[10] applications. Corpus linguistics and natural language processing methods have been applied with the purpose of frequency analysis of tokens and collocations using LeXimir, while the finite-state automation has been constructed in the Unitex software for the extraction of socio-political and culturological topics. The TreeCloud tool (Gambette and Véronis, 2009) has been used for visualising dominant topics in the corpus, within the text mining framework (serb. *kopanje po tekstu*, according

[7] The above English term is standardised in foreign literature, unlike in the articles and studies published in this area, in which numerous attempts of unique terminological determination are made (e.g. *nalaženje podataka na vebu* ("finding web data"), *struganje podataka* ("data scraping") or *grebanje veba* (web scraping), to name just a few). The aforementioned terms originate from the web articles "Nalaženje podataka na internetu" and "SEO optimizacija kroz rečnik najbitnijih termina".

[8] Codes are available at the author's GitHub repository (on-line).

[9] Slovo Majstor (on-line).

[10] LeXimir (on-line).

to Кешељ and Шипка (2008) or *iskopavanje iz teksta*[11]), which is based on discovering text patterns in unstructured data.

The paper consists of five sections. Basic features of Yu-rock, some relevant works in the field of computational processing of domestic and international song lyrics, as well as the research methods used in this paper were considered in the introductory part. The section 2 describes the process of automatic collection of material for the Yu-rock song lyrics corpus. The semi-automatic[12] preprocessing and corpus annotation methods were presented in the section 3. The experimental results of the application of the corpus are shown in the section 4, while the section 5 gives concluding remarks and guidelines for future work.

2 Data Collection via Web Scraping Methods

The narrower field of research proposed in this paper is concerned with the computational creation and analysis of the corpus of Yugoslav rock song lyrics during the era of two Yugoslav states – the Socialist Federal Republic of Yugoslavia (1945-1992) and the Federal Republic of Yugoslavia (1992-2003). As for the criteria for data collection, the corpus is composed of lyrics written in the former Serbo-Croatian language, which had been originally standardised by the Vienna Literary Agreement in 1850, but eventually split into Serbian, Croatian and Bosnian language, with the disintegration of Yugoslavia in 1991 (Hentschel, 2003). Accordingly, the lyrics in the mentioned three languages have been sought, whereas the lyrics written in other languages that had been in use in Yugoslavia – Macedonian, Slovene and minority languages – were excluded.

Regarding the web scraping method, it refers to the direct downloading of digitalised content with the aid of a specific computational tool. The core of this practice can be described as the automatisation of the comprehensive and relatively fast extraction and storage of web data. Web scraping is imposed as a far more efficient solution in comparison with the manual methods of copying and collecting each information unit individually. Additionally, the technique in question can represent the only possible solution for data extraction in case the "copy/paste" options are disabled in web browser.

[11] The latter term (lit. "excavation from the text") is taken from the presentation of the Prof. Dr. Cvetana Krstev (on-line).

[12] The semi-automatic method involves a combination of techniques for automatic and manual preprocessing and corpus annotations.

2.1 Description of the Data Source: the LyricWiki Website

LyricWiki[13] is a commercial music website which hosts lyrics of domestic and international artists. The website is searchable by name of artist, album, song, genre, and record label. Various lists are also available, such as the lists of the most popular albums from a certain year, according to the opinion of the editors of prominent music magazines, and lists of albums of film soundtracks, as well as plenty of other textual content related to the music field. As stated in the description of this website, LyricWiki is publicly available and is licensed to publish reliable lyrics.

The website's database stores more than 2,054,289 texts (data from June 15, 2019).[14] Since LyricWiki contains a relatively large number of songs in Serbian and former Serbo-Croatian language, at the initial stage of the research (while the website was still open for editing), the lyrics of the songs performed by selected musicians from the former Yugoslavia have been collected.[15] The additional reason for choosing this site for the scraping of textual data was the fact that lyrics could be extracted from it via the specific API, i.e. the `LyricWiki` module from the `lyricsmaster` library using the Python language, which will be discussed in detail in the next section.

2.2 Functionality of the lyricsmaster Library in the Python Language

Automatic collection of lyrics from the Internet represents an efficient and relatively commonly leveraged method that precedes the electronic corpus analysis. One of the representative examples of such a practice is the use of the `lyricsmaster` programming library, which is available on the PyPI's repository website[16] containing Python programming packages. Using this API, the lyrics stored in the databases of some of the most popular music websites (hereinafter referred to as *providers*), such as AZLyrics, Genius, etc., can be collected. On the same webpage, the usage of the aforementioned tool was demonstrated with the purpose of directly downloading and saving the lyrics of the late American rapper Tupac Shakur, which are available on LyricWiki. For the anonymisation of the user's IP address, there is an option for scraping the website via the Tor Proxy Server[17].

[13] LyricWiki (on-line).
[14] Statistics (on-line).
[15] The list of artists starts from the webpage Language/Serbian.
[16] lyricsmaster 2.8.1 (on-line).
[17] Tor Project (on-line).

The implementation of the aforementioned library has resulted in automatic extraction of textual content from various LyricWiki webpages, each of which was dedicated to some of the popular Yugoslav performers whose lyrics have been collected. Specifically, the web scraping function has been defined, and the `get_lyrics()` method of the class `lyricsmaster.providers.LyricWiki(tor_controller = None)`[18] has been used in order to retrieve the desired pages, from which only the lyrics were collected, and not the other content visible on the same webpages (e.g. header or sidebar information). An algorithm had been constructed which has collected the lyrics content by artist name in the following manner:

1. The LyricWiki website was firstly selected as the lyrics provider;
2. A list of thirty artists, whose lyrics were to be collected, was defined with the variable `izvodjaci` ("performers"). The artists' names were referenced exactly as they had been listed on the website (e.g. instead of `'Yu-Grupa'`, `'yu grupa'`, etc., entering the string value `'YU Grupa'` was only allowed). This also applied to cases in which two artists shared the same name (e.g. the Serbian and Finnish group Negative). The LyricWiki editors have made a difference between the mentioned groups by assigning the appropriate "RS" tag of the country of origin to the Serbian band, in order to explicitly refer to the group from the Republic of Serbia. With that in mind, the modified label `Negative (RS)`[19] was entered in the web scraping code;
3. The `corpus()` function was created whose arguments were the elements of the aforementioned list of artists, and for each of them it was attempted to collect the lyrics with the method `get_lyrics()`;
4. The `discography`, `album` and `song` objects were then created, while the introduction of the parameters `title` and `lyrics` resulted in accessing the album titles (`album.title`), song titles (`song.title`) and lyrics content (`song.lyrics`);
5. Afterwards, the collected corpus material was stored in the local computer memory using the `save()` method, which was applied to the `discography`, `album` i `song` objects.[20] The default absolute path was `/{user}/Documents/LyricsMaster/`, and the lyrics content were saved in the `/artist/album/` directory in the `song.txt` format.

[18] `get_lyrics` is the main method of the specified class (from the documentation on-line).

[19] The Finnish group did not receive any label, but only the name "Negative".

[20] `save()` is the method of three classes: `lyricsmaster.models.Discography`, `lyricsmaster.models.Album` and `lyricsmaster.models.Song` .

On the LyricWiki webpages with the lyrics of certain artists (e.g. YU Grupa), most of the song titles had been formatted as hyperlinks that pointed to the webpages with available lyrics. However, for several song titles the lyrics content had not been created at all (e.g. for the song „Čovek i Marsovac", i.e."A Man and a Martian"), despite the formal existence of the song titles, which was causing the algorithm to stop prematurely. In order to handle the exceptions, the `try` statement was introduced for the error detection in the place(s) where the problem(s) had occured. The `except` and `continue` statements, which eliminate the errors and allow the program to continue the started procedure, were also added. The mentioned errors are summarized in the Table 1.

2.3 Limitations of the lyricsmaster Library

It was not possible to collect lyrics with the `lyricsmaster` library for certain bands and solo artists (Riblja Čorba, Ekatarina Velika, Azra, Film, Gibonni, Đorđe Balašević, etc.), either due to the restricted access to data or the lack of lyrics by a specific artist on LyricWiki. Furthermore, at the stage of collecting the lyrics it was concluded that the corpus should contain female artists' lyrics, but for certain female artists whose musical style can be characterized as "rock" (Kaliopi, Slađana Milošević, etc.) their lyrics could not be automatically extracted from the given website. Likewise, some artists did not even exist in the database, and so as their lyrics (as was the case with Maja Odžaklijevska).

Song/Artist	Error	Cause	Solution
"Čovek i Marsovac"	`AttributeError`	Link unavailable	`try-except-continue`
`'Yu-Grupa'`	`TypeError`	Incorrect name	`'YU Grupa'`

Table 1: Problematic cases during the collection of lyrics.

In order to compensate for this deficiency, the alternative criteria have been applied: they refer to the subsequent selection of artists by expanding the genre's scope that will be covered by the corpus material. More precisely, the pop artists (such as Nina Badrić and Zana), whose songs have been influenced by rock music, were also included in the corpus. Other added

artists were Madame Piano[21] and Goran Karan[22]. The corpus proposed in this paper consists of songs of the artists listed in the Table 2:

Bajaga	Električni Orgazam	Neverne Bebe
Bajaga i Instruktori	Zabranjeno Pušenje	Negative
Bebi Dol	Zana	Nina Badrić
Bijelo Dugme	Idoli	Oktobar 1864
Van Gogh	Indexi	Partibrejkers
Galija	Josipa Lisac	Prljavo Kazalište
Goran Bregović	YU Grupa	Rani Mraz
Goran Karan	Kerber	Smak
Divlje Jagode	Madame Piano	Hari Mata Hari
Dino Merlin	Mirzino Jato	Haustor

Table 2: List of 30 artists in the corpus.

The oldest album in the corpus is *Naše doba* (1967) by the band Indexi; in contrast, the most recently published albums which are included in the corpus are the Divlje Jagode's album *Od neba do neba*, and *Collection* by Nina Badrić (both published in 2003).

2.4 Creating a Directory Tree for Storing the Corpus

By executing the web scraping code, a hierarchically organised data structure was created, that is, a directory tree with its root and branches. The resulting tree structure of the corpus can be represented in form of an output from the command line of an operating system, after running the code in the programming language Bash:[23]

[21] The Yu-rock expert, Petar Janjatović, included Madame Piano in his YU rock encyclopedia (Janjatović, 1998), owing to the impact that the aforementioned representative of jazz and world music sound had on the development of the Yugoslav music scene.

[22] Goran Karan is also not a representative example of a rock artist, but does not deviate too much from the target frame. Although Karan gained popularity by performing pop songs "with a Dalmatian tone", he had began his career as a rock singer (the information obtained from his biography on-line).

[23] Based on the Bash code provided on-line.

```
alias tree="find . -print | sed -e 's;[^/]*/;|____;g;s;____|; |;g'"
```

The Figure 1 depicts a partial view of the structure in question that was taken from the Terminal application included with the macOS operating system, after the initial navigation to the directory `LyricsMaster`, where the lyrics corpus was located.

Figure 1: Directory tree "LyricsMaster" from the command line.

As can be concluded from the Figure 1, the root directory carries the default name "LyricsMaster", within which the directories with the artists' names (e.g. Mirzino Jato or YU Grupa) were located. Then, for each artist, the album titles (*Šećer i med, YU Grupa, Rim 1994*, etc.) were listed. The albums contain lyrics in the .txt format which represent the terminal nodes in the tree structure (e.g. "Apsolutno tvoj", "Crni leptir", etc.). Among the mentioned data, the .DS_Store file also appeared, which was not of any significance for further corpus processing and analysis, and which was therefore eliminated at a later phase. After the web scraping procedure, the collected lyrics content in the .txt format, stored in separate directories, in the next stage of the corpus preparation was unified into an unique XML file and annotated.

3 Corpus Preprocessing

3.1 DTD Specification

Prior to the automatic generation of the corpus in the XML format, the *document type declaration* (abbr. DTD) had been created, which specified the logical structure of the XML document to which the document refers, and in relation to which its validity was checked. This step was also important because it represented a prerequisite for the further processing of the XML document, such as generating basic statistics from it in a clearly presented XHTML format using the XSLT processor, which will be discussed in the subsection 4.1. In the valid XML corpus document, for example, the DTD declares that the root element `<exYuPesme>` can have multiple authors, but that its attribute value is a fixed empty string.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="tabele-css-classes.xsl"?>
<!DOCTYPE exYuPesme [
<!ELEMENT exYuPesme (autor)+>
<!ATTLIST exYuPesme
xmlns CDATA #FIXED ''> ... ]
```

Figure 1: Document Type Declaration – excerpt.

3.2 Locating the Lyrics Using the os Module

The first step in the process of generating the XML document was the implementation of the `os` module using the `listdir()` method, from The Python Standard Library. When the method is called from the root "LyricsMaster" directory, it returns the directory list, i.e. the artists' names. In order to collect the names of the subdirectories (album titles of the given artist) and files (song titles on those albums), it was necessary to perform a string formatting in order to create the absolute paths using `format()` function. The final parts of the path, that is, a string of everything that comes after the last slash in the path argument (*base name*) represents the values of the target categories. In this process, two functions of the `os.path` module were used: with the `join()` function, the elements of the list of directories/files were concatenated to the ends of the current directory paths in order to access these elements. The second function, `isfile()`, was combined with the flow control statements `if` and `if not` in order to correctly determine whether a list of directories or files was generated. For illustration purposes, below are listed the paths based on which it was

possible to extract the name of the band Smak, their album title *Crna Dama* and the song title from the same album – "Daire":

```
../../LyricsMaster/Smak
../../LyricsMaster/Smak/Crna-Dama
../../LyricsMaster/Smak/Crna-Dama/Daire.txt
```

Then the lyrics content was loaded using the `open()` and `read()` functions, whereas the song structure of lyrics containing newlines was retained using the `split('\n')[:-1]` function. The purpose of the described procedure was to correctly allocate the data when annotating the file in the XML language.

3.3 Elimination of the Redundant Content

Certain lyrics from the corpus had been written exclusively in foreign languages. Since the subject of this paper was the processing of the corpus of lyrics in Serbian or Serbo-Croatian language, the manual removal of the mentioned songs or even the whole albums from the corpus directories was performed. In the present subsection, some representative examples of corpus preprocessing were listed. In particular, the lyrics written in Greek, Macedonian, Romanian, English, Portuguese and Polish language were excluded from the analysis. A specific case was the presence of multilingualism in the lyrics of Goran Bregović as a solo artist, who was also the author of the songs "Κέρνα μας", from the album *Alkohol: šljivovica & champagne*; "7/8 & 11/8", "Ederlezi", "TV screen", "Ausência" (album *Ederlezi*) and "To nie ptak" (*Kayah & Bregović*). On the album soundtrack from the film *Arizona Dream* all the songs were in English, so that album was also removed.

The lyrics which were attributed to an author by mistake were also excluded from the corpus. For example, among the lyrics on the Nina Badrić's LyricWiki webpage there was also the song "Ubila si del mene", which belongs to the Slovene boy band Game Over. Two songs ("Muistoja" and "Myrsky") by the homonymous band from Finland had also been incorrectly assigned to the Serbian group Smak from Kragujevac. Besides, on the album *Zašto ne volim sneg* by the band Smak there were several instrumental songs. The songs for which the lyrics content on LyricWiki had been missing (instead of the lyrics there were only the suspension points), as in the case of the song "Ne mogu da kapiram", by the group Partibrejkers, were also not taken into account. Moreover, there were the cases whereby the duplicates of lyrics appeared under different song titles. For example, the lyrics of the song

"Počasna salva" by the band Zabranjeno Pušenje were retained under that song title, while the duplicates with the false titles "Manijak", "Vuk" and "Ujka Sam" were deleted.

Particular attention was paid to selecting the songs from concert or compilation albums. Namely, it had been initially intended to immediately remove such directories, as it was expected that they would have contained the songs that had already existed in the corpus. However, the presence of certain number of unpublished songs on the concert albums was noticed (e.g. "Na vrhovima prstiju" from the album *Neka svemir čuje nemir* by the band Bajaga i Instruktori). From the greatest hits album *Collection* by Nina Badrić few unique songs were kept. Another type of duplicate was also found in form of some song covers, such as for the song "Tako ti je mala moja kad ljubi Bosanac", originally published by Bijelo Dugme in 1975 and covered by the group Zabranjeno Pušenje in 1998.

3.4 Annotating and Preprocessing the XML Corpus

`Yattag`[24] represents the Python programming library that can automatically insert HTML and XML tags while structuring the document. This API automatically inserts open and closed angle brackets, and each start-tag is followed by a end-tag. In this manner, the annotator can tag texts more quickly and easily, because the program reduces the possibility of reporting syntax errors. Using the module `indent`, the `Yattag` library also supports automatic indentation according to the general hierarchical structure of an XML document, and the size of the indented space can be modified. Since the annotation procedure was to be carried out on a large corpus, the objective was to exploit the functionality of the present library in order to annotate the document more efficiently, in accordance with the defined rules for marking up the lyrics content. the content of the texts. These rules concerned the inclusion of the XML elements' and attributes' tags for describing the corresponding parts of the corpus content using the following procedure:

- The root element was defined by the tag `<exYuPesme>`;
- The authors, i.e. performers, were defined by the element `<autor>`, whose attributes are `ime`, `brojAlbuma`[25], `pol`, `zanr`, `rodnoMesto` (songwriters were not included because they had not either been listed on the website);

[24] Yattag (on-line).

[25] That is, the number of albums included by the lyrics that form part of the corpus.

– The albums were defined by the `<album>` element, which has the attributes `naziv`, `godina` and `izdavac`;
– The songs were defined by the `<pesma>` element which has the `naslovPesme` attribute, while the tag of the `<li>` element was reserved for the verse lines.

Since the method `tag()` creates XML tags, only the desired tag names for marking up the elements and their attributes were forwarded as the method's arguments. For example, the album as an element has the attributes for the title, publication year and the record label, which was defined as follows: `with tag('album', naziv=album, godina="", izdavac=""):`. On the other hand, the method `text()` generated the text content that was not the name of the tag. For brevity of code, the `Doc` instance of the class `yattag.Doc` and the joint method `tagtext()`, which adds the content produced by the specified method to this instance (names of elements, attributes, and the plain text), were used. After defining all the necessary parameters, the `getvalue()` method was applied, in order to convert the whole content into a long character string. This way, the lyrics were transferred into a new XML file.

After the XML corpus had been experimentally generated, the special ampersand character (`&`) was noticed, which, as a rule, was replaced by the escape sequence character (`&`). This was performed in order not to interpret the ampersand as the beginning of the entity reference and to fully comply with the principles of proper XML structuring. For that reason, the `escape()` function of the module `xml.sax.saxutils` was inserted into the code for annotating the corpus. Although the apostrophe symbol (`'`) often appears in the corpus, it was not replaced by its XML equivalent `&apos;` in the process of automatic character substitution, which did not hinder the process of generating a well-formed XML document. For the sake of greater transparency, the hyphens present in the titles of the songs and albums that had remained during the web scraping procedure, were automatically removed (e.g. the initial song title "Da-Sam-Pekar" was changed to "Da Sam Pekar").

As for the manual annotation of the corpus, the values "pop", "rock" and "world music"[26] of the attribute `zanr` were appended to the element `<autor>`. Besides the genre, the values of the attribute `rodnoMesto` which refer to the artist's birthplace and `izdacac` for the record label that published the album, were also added. An example of the semi-automatically annotated lyrics from

[26] According to the genre classification available on the website Discogs.

the the album *Koncert kod Hajdučke česme* by Bijelo Dugme, can be seen in the Figure 2.

```
<autor ime="Bijelo Dugme" brojAlbuma="13" pol="Grupa" zanr="Rok" rodnoMesto="Sarajevo">
    <album naziv="Koncert kod hajdučke česme" godina="1977" izdavac="Jugoton">
        <pesma naslovPesme="Da Sam Pekar">
            <li>Da sam pekar, mala moja</li>
            <li>Ne znam bi l'□ me htjela</li>
            <li>Kad bi noću bila sama</li>
            <li>Zemičke bi jela</li>
```

Figure 2: Excerpt from the annotated corpus in the oXygen XML Editor software.

3.5 Automatic diacritic restoration – LeXimir software

The corpus normalisation and its preparation for the computational analysis does not represent a trivial task whatsoever, and, as a rule of thumb, it contributes to generating more informative results in comparison with the computational analysis of the non-preprocessed text. The collected lyrics in the Yu-corpus were rather uneven in terms of the use of the writing systems: most of the lyrics were written in Latin script, and there were quite a few lyrics in which the special Latin letters (č, ć, ž, đ, š) did not contain the necessary diacritical marks. Simultaneously, fewer songs were originally written in Cyrillic. The initial idea had been that the whole corpus be translated into Cyrillic; this idea was later rejected because the lyrics in the Serbian language also contained some excerpts in foreign languages (e.g. "la musique c'est fantastiqu / prepare la revolution / et la femme est tres jolie / tre jolie comme un bonbon").[27] It was therefore decided that during the transliteration process the lyrics written in the degraded Latin and Cyrillic alphabet be automatically converted into the standardised Serbian Latin alphabet containing diacritical marks. The transliterated corpus lyrics in the .txt format were used for extracting lexical units (see the subsection 4.2).

In order to select the most suitable tool, the diacritic restoration was performed in the Slovo Majstor and LeXimir applications. The second software uses electronic morphological dictionaries (with valid word forms), local grammars and the Corpus of Contemporary Serbian Language

[27] According to the original lyrics, with typographical and spelling errors.

(Krstev et al., 2018). After evaluating both methods, it was concluded that Slovo Majstor solidly solved the problem of converting the degraded Latin alphabet into the standardised one, with the remark that the application incorrectly added diacritical marks to some words, which violated the word semantics. Thus, from the verb form "isecka" the word "isečka" was incorrectly created. The favourable aspect of this application was the ability to subsequently correct the incorrectly transliterated words.

Conversely, the LeXimir tool, which exploits the functionalities of Unitex corpus processing software, left the previously mentioned word in the correct form. In the Figure 3, the potential candidates for selecting an adequate form were the word "isecka", which is an inflected form of the noun "isečak" in the genitive singular case, and "isecka", that is, third-person singular of the aorist of the verb "iseckati". The diacritic restoration algorithm based on the left and right textual context of the potential word for the lexical correction performed the *disambiguation* (ambiguity resolution). The second option was selected, where the pronoun "ga" can co-occur with the verb "isecka", and not with the inflected form of the noun "isečka", since the other construction would be grammatically incorrect. Nevertheless, manual evaluation is to be performed after the implementation of the described tool as well, because in some words (e.g. in the words "oci" in the context "Trljam oci sanjive"), no diacritical marks were added where they had been supposed to exist.

Potencijalni kandidati	Levi kontekst	Ispr.	Desni kontekst
*7(uže(72)_uze(61))	********	uze	********
*7(Isečka(5)_Isecka(1))	s I ujeo kuvara Kuvar uze, uze nož	Isecka	ga na dva, na tri komada Došli, do
*7(oči(818)_oci(24))	žurno Kafu pijem, nestajem Trljam	oci	sanjive Da mi ne bi zaspale Vrata

Figure 3: Diacritic restoration in the LeXimir tool.

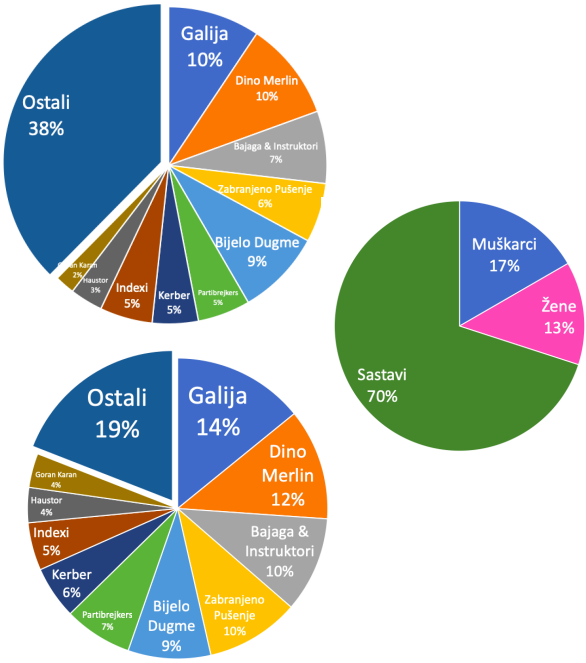
4 Computational Corpus Analysis

4.1 Corpus Statistics

The DTD specification enabled submitting queries by navigating over the given document using the XSLT language for transforming an XML document into other formats. For instance, the following XSLT expression selects the value of the attribute `ime` for each element `<autor>` in the document, which generates a list of all artists in the corpus:

```
<xsl:for-each select="//autor">
        <xsl:value-of select="@ime" />
</xsl:for-each>
```

Data from the XML document were transformed into the XHTML table from the Figure 4, where it can be observed the total number of artists, songs and albums in the corpus. The pie charts from the same figure illustrate (clockwise, top-down): the artists' percentage in the corpus by the number of albums, the percentage of male, female and group artists, as well as the percentage by the number of songs ("ostali" means "others").



Statistika

Ukupno: 30	*Ukupno: 953*	*Ukupno: 147*
Autori	**Pesme**	**Albumi**
Galija	135	14
Dino Merlin	114	15
Bajaga Instruktori	98	11
Zabranjeno Pusenje	96	9
Bijelo Dugme	85	13
Partibrejkers	70	8
Kerber	54	7
Indexi	49	8
Haustor	36	5
Goran Karan	35	3
Prljavo Kazaliste	25	6
Negative	22	2
Hari Mata Hari	22	7
Smak	21	6
Nina Badric	20	5
Idoli	15	1
YU Grupa	14	7
Rani Mraz	13	2
Divlje Jagode	6	4
Madame Piano	5	1
Goran Bregovic	4	2
Van Gogh	2	2
Josipa Lisac	2	2
Bajaga	2	1
Zana	2	2
Oktobar 1864	2	1
Mirzino Jato	1	1
Neverne Bebe	1	1
Elektricni Orgazam	1	1
Bebi Dol	1	1

Figure 4: Yu-corpus statistics.

LeXimir also produced basic statistics of the Yu-corpus. After the diacritic restoration, the corpus contains 248,807 tokens, 16,964 unique lexical

units, 116,972 simple forms (16,909 different) and 268 numbers (10 different). LeXimir also provides support for displaying inflected forms of tokens, lemmatised forms, parts of speech, words' and collocations' frequencies. The results of the frequency analysis can be filtered by the part of speech. Based on the exported .xlsx file it can be determined, for instance, what are the most frequent nouns or collocations whose headword is a noun. The Figure 5 below shows the partial display of the results which reveal the most frequent noun tokens, among which the words *dan, noć, ljubav, srce*, etc., appear. The Figure 6 provides an overview of certain collocations related to war (*svetski rat, vojnu muziku, ratne filmove*).

Oblik	Lema	POS	Freq.
dan	dan	N	299
Al	Al	N	231
noć	noć	N	228
do	do	N	224
ljubav	ljubav	N	201
srce	srce	N	196
život	život	N	195
put	put	N	184
meni	mena	N	164
meni	meni	N	164
kraj	kraj	N	155
noći	noć	N	147
biti	bit	N	132
bila	bilo	N	124
grad	grad	N	120

Figure 5: Tokens.

svetski rat	svetski rat	N	3
novi svet	Novi svet	N	3
noćne ptice	noćna ptica	N	3
tam-tam	tam-tam	N	3
prošlog vremena	prošlo vreme	N	2
vojnu muziku	vojna muzika	N	2
železnička stanica	železnička stanica	N	2
crno grožđe	crno grožđe	N	2
sunčev zrak	sunčev zrak	N	2
malog medveda	Mali Medved	N	2
zlatne medalje	zlatna medalja	N	2
morske obale	morska obala	N	2
svetla budućnost	svetla budućnost	N	2
ratne filmove	ratni film	N	2

Figure 6: Collocations.

4.2 Extracting topics from the corpus

It is known that the socio-political and culturological topics were present in the Yugo-rock lyrics, thus the Unitex software was used for constructing the graph of finite-state automation for recognising the indicators of the above topics. Regular expressions and lexical masks served for the retrieval of lexical forms in the form of abbreviations[28] composed of:

- capital letters, hyphens and a sequence of numbers (e.g. *B-52*);
- sequence of capital letters + full point repeating at least two times (*K.P.*);

[28] See the webpage Skraćenice i objašnjenja.

– sequence of capital letters + space repeating at least two times (*ES EF ER JOT*);
– sequence ranging from two to five capital letters (*CZ, CIA*).

Among them the names of state authorities (e.g. *AVNOJ*), sports clubs (*PFC*), the name of broadcasting companies (*BBC*), etc. came to the fore. The graph from the Figure 7, in addition to the abbreviated names presented in the concordance result in the Figure 8, also recognised the tokens: *CZ, ES EF ER JOT, FK, TV, JRT, K.P., KGB , KK, MUP, O.K., PC, PFC, SUP* and *TAS*.

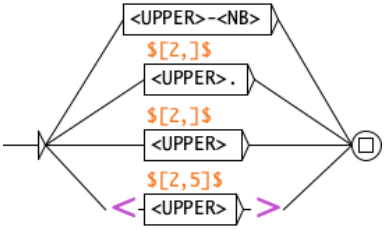


Figure 7: Graph that recognises the abbreviations.

Od istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
nebu je pisalo B-52 I veliki buketi od
nebu je pisalo B-52 I veliki buketi od
šao Haše!{S} I BBC na mome radiju, Osl
šao Haše!{S} I BBC na mome radiju, Osl
niskim terenom BMW-a kešom plaćenog, o
KGB zaspala je CIA odmara se naša mili
, penjao se na CK i pevao pesnu protiv
iti stan Preko CNN-a gledao sam Mufu S

Figure 8: Concordance excerpt.

In order to determine whether the socially engaged topics were statistically significant in the corpus, a visual overview of predominant topics in the form of a *tree cloud* had been generated using the TreeCloud tool, with the remark that the analysis had been performed on non-lemmatised forms. This program combines the word cloud graphical representation of data with a tree structure, so that, apart from the word frequencies, it is also possible to observe the way of clustering the tokens based on their distance in this structure (see the Figure 9). The application also includes a built-in stop words list for Serbian language, which was supplemented with some words for the sake of better visualisation of the given corpus. A visual representation was created, in which one can distinguish the clusters indicating feelings, expressed by the words *tuga, ljubav/-i, bol, srce/-a*. Urban topics were represented by the words *grad/-u, ulice*, body parts were associated with the words *ruke, lice, oči, srce/-a/-u*. The motive of time is also frequent in the corpus (*život, vr(ij)eme, godina/-e*). According to this visualisation, socially engaged topics are not sufficiently present in the corpus.

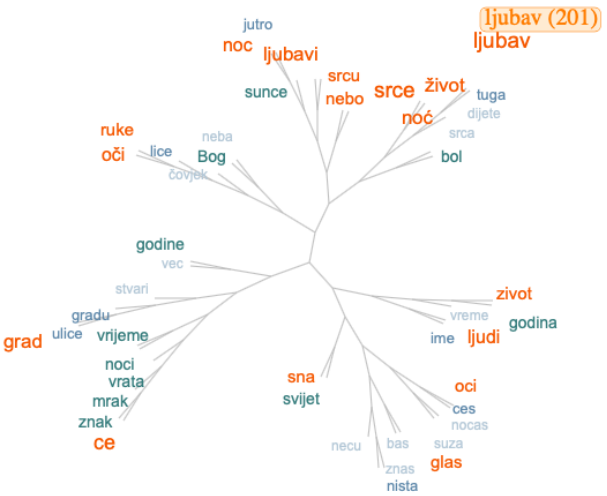


Figure 9: Tree cloud of the whole corpus.

5 Concluding Remarks

The paper discussed the project of electronic creation and processing of the Yugoslav rock song lyrics from 1967 to 2003. Automatic data collection from the LyricWiki website was performed using the `lyricsmaster` library in the Python programming language, and the preprocessed corpus was automatically annotated in compliance with the XML syntax rules, using the `yattag` tool, along with the manual adding of some attribute values. Automatic diacritic restoration was also carried out. XSL transformation of the corpus into XHTML format was also shown, as well as the extraction of socio-political and culturological topics in the Unitex software and the visualisation of the prevailing topics with the TreeCloud tool.

Further work would include the corpus evaluation with the aim of correcting the spelling and typographical errors (e.g. *Dunav* instead of *dunav*, *Avale* instead of *Aavale* etc.). Also, the implementation of electronic morphological dictionaries of named entities is planned, in order to extract the names of musicians, politicians, athletes and other celebrities who had exerted a significant impact on the Yugoslav society.

References

Bertin-Mahieux, Thierry, Daniel PW Ellis, Brian Whitman and Paul Lamere. "The million song dataset". In *Proceedings of the 12th International Conference on Music Information Retrieval*, 591–596. 2011. Accessed: 22/10/2019, http://ismir2011.ismir.net/papers/OS6-1.pdf.

Cooper, Laura E. and B. L. Cooper. "The pendulum of cultural imperialism: Popular music interchanges between the United States and Britain, 1943-1967". *Journal of Popular Culture* Vol. 27, no. 3 (1993): 61–78. Accessed: 22/10/2019, https://sci-hub.tw/10.1111/j.0022-3840.1993.00061.x#.

Falk, Johanna. "We will rock you : A diachronic corpus-based analysis of linguistic features in rock lyrics". Bachelor's thesis, Linnaeus University, Department of Languages, 2013. Accessed: 22/10/2019, http://www.diva-portal.org/smash/get/diva2:605003/FULLTEXT02.pdf.

Gambette, Philippe and Jean Véronis. "Visualising a Text with a Tree Cloud", *IFCS'09: International Federation of Classification Societies Conference* (2009): 561–569. Accessed: 22/10/2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643/file/2009GambetteVeronis.pdf

Haslam, Thomas J. "Mapping the Great Divide in the Lyrics of Leonard Cohen". *Rupkatha Journal on Interdisciplinary Studies in Humanities* Vol. 9, no. 1 (2017): 1–10, Accessed: 22/10/2019, http://rupkatha.com/V9/n1/v9n1s01.pdf

Hentschel, Elke. "The expression of gender in Serbian". In *Gender across languages: The linguistic representation of women and men*, Vol. 3, John Benjamins Publishing Company, 287–309, 2003. Accessed: 22/10/2019, https://epdf.tips/gender-across-languages-volume-iii-the-linguistic-representation-of-women-and-me.html

Janjatović, Petar. *Ilustrovana YU rock enciklopedija: 1960-1997*, Geopoetika, 1998. Accessed: 22/10/2019, https://monoskop.org/images/c/ca/Janjatovic_Petar_Ilustrovana_YU_Rock_Enciklopedija_1960-1997.pdf

Kreyer, Rolf and Joybrato Mukherjee. "The style of pop song lyrics: A corpus-linguistic pilot study". *Anglia-Zeitschrift für englische Philologie* Vol. 125, no. 1 (2007): 31–58, Accessed: 22/10/2019, https://doi.org/10.1515/ANGL.2007.31

Krstev, Cvetana, Ranka Stanković and Duško Vitas. "Knowledge and Rule-Based Diacritic Restoration in Serbian". In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria* (2018): 41–51, Accessed: 22/10/2019, https:

//www.researchgate.net/publication/328416358_Knowledge_and_Rule-Based_Diacritic_Restoration_in_Serbian

Lightman, Erin J., Philip M. McCarthy, David F. Dufty and Danielle S. McNamara. "Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters". In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29, 2007. Accessed: 22/10/2019, https://cloudfront.escholarship.org/dist/prd/content/qt0dh4553j/qt0dh4553j.pdf.

Lukic, Alen. "A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics", Tech report, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, s.d. Accessed: 22/10/2019, http://alenlukic.com/assets/docs/lyric_topic_modeling.pdf

Mahedero, Jose P.G. Álvaro Martínez, Pedro Cano, Markus Koppenberger and Fabien Gouyon. "Natural language processing of lyrics". In *Proceedings of the 13th annual ACM international conference on Multimedia*, 475–478. 2005. Accessed: 22/10/2019, https://www.researchgate.net/profile/Pedro_Cano5/publication/221573745_Natural_language_processing_of_lyrics/links/00b7d52826f623edfb000000.pdf

McEnery, Tony and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012. Accessed: 22/10/2019, http://gen.lib.rus.ec/book/index.php?md5=33c1c5b6d73ea816dfb2a034f73bb176

Petrie, Keith J., James W. Pennebaker and Borge Sivertsen. "Things We Said Today: A Linguistic Analysis of the Beatles". *Psychology of Aesthetics, Creativity, and the Arts* Vol. 2, no. 4 (2008): 197. Accessed: 22/10/2019, https://www.uvm.edu/pdodds/files/papers/others/2008/petrie2008a.pdf.

Stein, Daniel. "Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes". In *LREC '2012 Workshop: LRE-Rel, Language Resources and Evaluation for Religious Texts*, 88–93. 2012. Accessed: 22/10/2019, https://www.academia.edu/26035335/Linguistic_and_Semantic_Annotation_in_Religious_Memento_mori_Literature

Taina, Jesse et al.. "Keywords in heavy metal lyrics: A Data-Driven Corpus Study into the Lyrics of Five Heavy Metal Subgenres", 2014. Accessed: 22/10/2019, https://helda.helsinki.fi/bitstream/handle/10138/136524/keywords.pdf?sequence=1.

Taylor, Charlotte. "What is corpus linguistics? What the data says". *ICAME journal* Vol. 32 (2008): 179–200. Accessed: 22/10/2019, http://sro.sussex.ac.uk/id/eprint/53389/1/what_is_corpus_linguistics.pdf

Whissell, Cynthia. "Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon". *Computers and the Humanities* Vol. 30, no. 3 (1996): 257–265. Accessed: 22/10/2019, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.7171&rep=rep1&type=pdf.

Zhang, Shuo, Rafael Caro Repetto and Xavier Serra. "Understanding the expressive functions of jingju metrical patterns through lyrics text mining". In *18th International Society for Music Information Retrieval Conference*, 397–403. 2017. Accessed: 22/10/2019, https://repositori.upf.edu/bitstream/handle/10230/32652/Zhang_ISMIR2017_unde.pdf?sequence=1&isAllowed=y

Zörnig, Peter, Emmerich Kelih and Ladislav Fuks. "Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis". *Glottotheory* Vol. 7, no. 1 (2016): 41–66. Accessed: 22/10/2019, http://homepage.univie.ac.at/emmerich.kelih/wp-content/uploads/2016_Zoernig_Kelih_Fuks_www.pdf.

Арсенијевић, Александра, Милена Обрадовић and Михаило Шкорић. "Израда мултимедијалног документа „YU рок сцена"". *INFOtheca – Journal for Digital Humanities* Vol. 16, no. 1-2a (2016): 113-129. Accessed: 22/10/2019, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2016.16.1_2.6_sr

Божиловић, Никола. "Култура сећања и југословенски рокенрол". *Култура* no. 152 (2016): 257–280. Accessed: 22/10/2019, https://scindeks-clanci.ceon.rs/data/pdf/0023-5164/2016/0023-51641652257B.pdf

Гајић, Златомир. "Рок поезија Бранимира Штулића и њени медијски одјеци". PhD thesis, Faculty of Philosophy, University of Novi Sad, 2018. Accessed: 22/10/2019, http://nardus.mpn.gov.rs/bitstream/handle/123456789/9926/Disertacija17602.pdf?sequence=1&isAllowed=y

Кешељ, Владо and Данко Шипка. "Приступ изградњи стемера и лематизатора за језике с богатом флексијом и оскудним ресурсима заснован на обухватању суфикса". *INFOtheca – Journal for Digital Humanities* Vol. 9, no. 1-2 (2008): 21–31. Accessed: 22/10/2019, http://infoteka.bg.ac.rs/pdf/Srp/2008/04%20Vlado-Danko_Stemeri.pdf

Раковић, Александар. “Бит мода, рокенрол и генерацијски сукоб у Југославији 1965-1967”. *Етноантрополошки проблеми* Vol. 6, no. 3 (2011): 745–762. Accessed: 22/10/2019, https://www.eap-iea.org/index.php/eap/article/view/604/594

Раковић, Александар. “Рокенрол у Социјалистичкој Југославији: од забаве градске омладине до националне културе”. In *Сан о граду : зборник радова*, 427–439. The Andrić Institute, 2018. Accessed: 22/10/2019, http://doi.fil.bg.ac.rs/pdf/eb_book/2018/ai_san_o_gradu/ai_san_o_gradu-2018-ch18.pdf

Textometric methods and the TXM platform for corpus analysis and visual presentation

UDC 811.163.41'322
DOI 10.18485/infotheca.2019.19.1.2

Jelena Jaćimović
jelena.jacimovic@stomf.bg.ac.rs

University of Belgrade
School of Dental Medicine
Belgrade, Serbia

ABSTRACT: Textometric approach has long been applied as a useful method for corpus analysis in various fields of humanities and social sciences. Textometry allows the non-linear quantitative and qualitative study of digital corpora, combining lexicometric and statistical research with developed corpus technologies. In this paper, the current version of the srpELTeC corpus was analyzed within the TXM program environment to illustrate the possibilities of the textometric approach and visual presentation of the obtained results.

KEYWORDS: textometry, digital corpora, Serbian language, TXM, srpELTeC.

PAPER SUBMITTED: 29 June 2019
PAPER ACCEPTED: 6 September 2019

1 Introduction

Digital corpora serve as valuable and practical sources of empirical data necessary for linguistics and other humanities and social sciences research. With the development of the digital era and the achievements of language technology use, the traditional way of accessing the text changes, opening new possibilities for analyzing large amounts of textual data using statistical methods. Unlike conventional linear reading (close reading), distant reading enables an understanding of literature by collecting and analyzing large amounts of data. Textometry stands out as one of the disciplines which allows a different way of "reading" the texts. Textometry enables the non-linear quantitative and qualitative study of digital corpora, combining lexicometric and statistical research with developed corpus technologies.

1.1 Textometry

The beginnings of textometric research relate to France and the work of Pierre Giraud (1954; 1959) and Charles Muller (1973), who dealt with problems and methods of linguistic statistics. The methods developed by Jean-Paul Banzécri and his colleagues and students, also applied to linguistic data (Benzécri, 1973), are adopted and implemented in textometry. The methodological basis of textometry could be also found in the works of Ludovic Lebart and André Salem (Lebart and Salem, 1988, 1994). In addition to the methods adopted, textometry has developed new statistical models to discover important features of textual data, such as contextual „attractiveness" of the words, the linearity and internal text structure, intertextual contrasts or the indicators of lexical evolution (characteristic period of a word usage, detection of the significant usage disruptions). The textometry analysis results give a synthetic, selective, and suggestive overview of the re-organized text, seen now through the hierarchical lists, visual maps, re-grouping, and text enhancements. The new way of accessing versatile, controlled textual data and a new way of „reading" the text based on the data that were previously not available, highlights the heuristic power of statistical tools in the analysis of literary texts.

A textometric approach implies that each text has its internal structure, which would be difficult to analyze manually. Facilitated by computer tools and created hypertext links, textometry based on numerical indicators simultaneously provides a general synthetic view of the text, as well as the possibility of local insight into the context. This text „closeness" during analysis, as well as a balanced approach to both the general and the local in the text, opens a range of hermeneutical questions and reveals a linguistic reality that is a highly significant representing valuable observational field. However, it should be kept in mind that textometry is a process that provides results and the identification of patterns and trends that would otherwise remain hidden due to large amounts of data, but that the interpretation of the obtained results and its validity depend on the experts and the system used for textometric analysis.

Beside textometry, other disciplines also use quantitative approaches for the analysis of textual data. Information retrieval (IR) deals with finding robust methods for managing large quantities of texts too. Nevertheless, IR focuses on finding and linking certain documents and information discovered in them. In contrast to the field of IR, textometry applies to a closed, stable corpus of texts. Another area that can compare with textometry in a cer-

tain sense is Latent semantic analysis (LSA). LSA also uses mathematical methods for the text analysis, but to investigate some out of the text fields, such as language and other cognitive functions. Quantitative methods that are known and applied in textometry are also used, for example, in the field of Text mining (TM). Unlike TM's basic idea to find and extract remarkably valuable information, the textometric approach focuses on the text itself and the discovery of linguistic trends and rules, their text realizations and changes. Furthermore, the subjects of textometric analysis are well-known corpora, documents, and objects. Natural language processing (NLP) also uses statistics for the system building and recognizing the linguistic units of the text. Although the objectives of these two areas differ, arguably they complement one another. For example, corpora can be used in the NLP to set up a system designed to identify specific entities or to build recognition rules (such as term extraction or morphological and syntax tagging), which can be of great importance in the later textometric analysis' process. On the other hand, the textometric approach to the corpus research provides the possibility of recognizing some entities or the text specificities, as well as text tagging, useful for the NLP tools development.

1.2 The TXM programming environment

The quantitative approach to researching textual corpus elements is not new, but it is significantly simplified and improved using existing tools. The programs designed to analyze large corpora do not lead to the discovery of new language information but offer a novel perspective on the perception of the already known (Hunston, 2002). When analyzing large corpora, it is essential to allow users to run multiple different queries and navigate through text in an easy-to-use environment. Several software solutions, which are free to use and have developed graphical user interface (Pincemin, 2018), allow text data analysis using the textometry primary functions (such as the concordance view, the *specificity score*, or correspondence factor analysis discussed later).

Based on textometry, a methodology that allows quantitative and qualitative analysis of textual corpora, TXM (Heiden et al., 2010; Heiden, 2010) is an open-source program,[1] widely used in research in various fields of social sciences (history, literature, geography, linguistics, sociology, political sci-

[1] Desktop installations for Windows, Linux, and Mac OS X are available, as well as a web platform.

ences). The graphical user environment of the TXM uses the CQP[2] (Corpus Query Processor) browser and the R[3] statistical package. The TXM allows the study of abundance of any language materials, a large number of input text formats, the use of various tools for processing natural language inputs (such as automated lemmatizers). This program enables the construction of a sub-corpora or parts based on metadata (date, author, genre, etc.) or corpus structural units (like text, chapter, paragraph), querying (using the CQP browser), and more complex query results processing based on quantitative methods, as well as the export of results in a tabular or graphical form.

Digital corpora possible to analyze within this environment are written texts, transcripts (synchronized with original audio or video), and parallel corpora. The TXM allows the import of texts encoded under certain conventions, such as texts in the recommended UTF-8 (TXT format) or XML[4] (eXtensible Markup Language) documents encoded following TEI[5] (Text Encoding Initiative) instructions (XML or XML-TEI format). Hence, it is possible to select, within the TXM, a representation level of corpus texts that is more or less rich and therefore more or less demanding for preparation. The plain text representation offers some essential analysis options. On the other hand, due to the ampler text representation and its internal structure, XML-TEI texts can be explored in far more detail (Lavrentiev et al., 2013).

During the import, TXM generates a customized version of the TEI data model - an XML-TXM format based on source texts and formats, used as a base model for all analyzes. This corpus representation implies the existence of specific corpus units, such as textual, structural, and lexical units. Textual units are **texts** that the corpus comprises (such as books, articles, interviews) and they can have their attributes or metadata (like author, title, date, genre). Then, each of the texts can have a specific structure and several internal **structural units** (for instance, chapters, passages, administrative speech) that may have particular properties or attributes (such as an address, number). Lastly, **lexical units** are defined, because each text is composed of a series of words that can have specific properties, such as graphic form, lemma, or grammatical category. Metadata, like all XML text elements, is considered within the TXM as a structural unit. Thus, the pos-

[2] CQP (on-line)
[3] R (on-line)
[4] XML (on-line)
[5] TEI (on-line)

sibilities of text analysis depend on its representation. Existing text/corpus annotations, along with structural and lexical units, are used to create a sub-corpus or corpus parts to compare them and search. Therefore, from the corpus research standpoint, it is necessary to define in more detail the units used for the analysis. The TXM builds an HTML format for each corpus textual unit, providing the possibility of returning to the text at any analysis phase.

Although TXM text import modules provide automatic morphological and syntactic tagging, as well as lemmatization of texts using the TreeTagger (Schmid, 1994), it is possible to preprocess corpus data beyond the platform using other NLP tools. There is no standard representation of the results of these tools, and only the standards used in practice apply. The results of NLP tools TXM recognizes as annotations added to the XML-TEI text representation.

In addition to built XML-TXM format, the CWB[6] (Corpus Workbench) format is also generated, applied to search the corpora using queries expressed by CQP syntax. A description of the CWB environment and regular expressions, used by Corpus Query Language (CQL), can be found in (Utvić, 2014; Evert and Hardie, 2011).

The TXM environment primarily enables qualitative text analysis through generated frequency lists, concordances, or the HTML text edition. Any combination of the properties of defined text units can be used to query and display the contexts in which those units appear. On the other hand, statistical models, counting the properties of lexical units, permit quantitative analysis, i.e. analysis of their corpora distribution (factor analysis, cluster analysis), their remarkably high or low representation in certain parts (specificity analysis), or the analysis of the lexical attraction between words (co-occurrence analysis). Each result of the analysis can be exported for further examination and editing, using another tool, in tabular or graphical form.

This paper aims to present the current version of the srpELTeC corpus[7] and to illustrate the possibilities of the textometric approach and the application of the TXM tools for analysis and visual presentation of the results. The conducted analysis of the Serbian novel corpus from the late 19th and early 20th centuries should highlight the potential of textometry and bring

[6] CWB (on-line)
[7] srpELTeC (on-line)

it closer to researchers from the various scientific fields involved in corpus analysis.

2 The methodology of the srpELTeC corpus textometric analysis

2.1 The srpELTeC corpus

The corpus used for this paper is called the srpELTeC corpus. The main motive for the creation of this corpus is its inclusion in a multilingual European Literary Text Collection, which should contain 100 novels for each of the languages included in the COST Action *Distant Reading for European Literary History*, published between 1850-1920 that have expired copyright.[8]

The Serbian corpus, unlike many other European languages involved in this action, is produced from the very beginning. Most Serbian novels from this period were not digitized or properly digitized, especially since the first editions of many novels were hard to obtain. Serbian literature, and especially the Serbian novel, can by no means be compared in scope with the literary „production" of the major European languages, such as French, English, or German. Therefore, the novel selection and finding printed copies were extremely demanding. Transformation into a machine-readable form involved scanning and optical character recognition (OCR). OCR errors were automatically corrected using a specialized tool based on the Serbian morphological dictionary (Krstev, 2008). A large number of volunteers were engaged in manual correction of the remaining errors and structural annotation markup.[9] At this stage, following the requirements of the COST Action, the metadata was also prepared to be used for later corpus analysis.

It is well-known that corpus size, its representativeness, and balance should be taken into account when designing a corpus (O'Keeffe and McCarthy, 2010). The preparation of several novels published in the Serbian

[8] The ELTeC collection should contain the first editions of literary texts (novels) from a distinct period and written in several languages. To be included in one of the ELTeC sub-collections, the text must have been first published as a book (minimum length of 10.000 words) in a European country between 1850 and 1920. Other novel selection criteria primarily concern the author's gender and the canon. Each ELTeC sub-collection should contain at least 10% to 50% of texts written by female authors, as well as at least 30% of both prestigious (highly canonized, reprinted more than once) and unknown (not or once reprinted) novels.

[9] Manual correction (on-line)

language in the period 1850-1920 is in progress, and for the time being, 21 works have been included in the ELTeC corpus, digitized until the writing of this paper (Table 1). The reason for the selection of these works, therefore, is neither an aesthetic nor a thematic nature. Since the SerbianELTeC corpus currently does not include all the novels published in a given period, it cannot be said to be representative or balanced. Besides, for example, most of the works included were written by male authors. However, this paper aims to demonstrate the implementation of textometric analysis using the TXM tool for which the created Serbian literature corpus from the late 19th and early 20th centuries can be a good source. Furthermore, this specialized corpus contains a collection of texts of exceptional significance that are not an example of a modern language and in which, besides well-known authors and their works, there are those whose work brings the beginnings of a modern novel structure or those about which is written insufficiently in the history of Serbian literature. For example, the corpus includes novels of Milutin Uskoković, whom critics and historians of literature considered as the originator of Belgrade's, urban style, but also a novel by little-known author Dragomir Šišković. Moreover, the first Serbian science-fiction novel *Jedna ugašena zvezda* by Lazar Komarčić is part of the srpELTeC corpus too, as well as the novel *Babadevojka* by Draga Gavrilović, the first female author who wrote the novel in the Serbian patriarchal society of the time. The current version of the srpELTeC corpus is available in the ELTeC collection.[10]

The texts of the srpELTeC corpus are encoded in XML format, very useful for later analysis, following the TEI guidelines. The header of the TEI document contains bibliographic information about the electronic and original version of the novel, as well as information about the persons responsible for the particular phases of creating and updating the electronic version. Documents are structurally annotated, containing information on the logical structure of the text. In addition to the recommended TEI elements and attributes for structural text annotation (like header, text, body, unit of text - chapter, title and subheadings, paragraph, quotation, words or sentences written in the language other than the language of the text), metadata in the form of a CSV document also includes supplementary information about the gender of the author and the publication type (like novel, story or short prose).

[10] ELTeC collection (on-line)

Author	Publication	Year	Length (w)
Gavrilović, Draga	Babadevojka	1887	23.858
Gavrilović, Andra	Prve žrtve	1893	44.929
Kostić, Tadija	Gospoda seljaci	1896	39.349
Mijatović, Čedomilj	Rajko od Rasine	1892	50.305
	Ikonija, vezirova majka	1891	28.332
Milićević, Milan	Deset para	1881	12.365
	Jurumusa i Fatima	1879	21.947
Stanković, Borisav	Uvela ruža	1899	12.748
	Pokojnikova žena	1902	12.701
Šišković, Dragomir	Jedan od mnogih - roman iz prestoničkog života	1920	21.676
Uskoković, Milutin	Potrošene reči	1911	14.580
	Došljaci	1910	97.467
	Čedomir Ilić	1914	65.073
Dimitrijević, Jelena	Nove	1912	116.782
Ilić, Dragutin	Hadži Đera	1904	65.554
Janković, Milica	Kaluđer iz Rusije*	1919	8.279
Komarčić, Lazar	Dragocena ogrlica	1880	65.160
	Jedna ugašena zvezda	1902	58.334
	Prosioci	1905	28.327
Nušić, Branislav	Opštinsko dete	1902	77.994
Sekulić, Isidora	Đakon Bogorodičine crkve	1919	62.414

* This publication will not be included in the srpELTeC corpus because of its length

Table 1. Literary works included in the srpELTeC corpus used for textometric analysis

The TXM environment recognizes defined metadata as a new structural element `text`, represented by the following attributes: `author`, `title`, `date`, `gender` and `type`. Thus, the distribution of the prepared texts was determined by the author's name, the title, year of publication, the author's gender, and the publication type. Metadata was used in the TXM environment to split the corpus into parts, create a sub-corpora, and for text search.

For the analysis of the srpELTeC corpus, a collection of texts in XML format was imported into the TXM environment using the XML-TEI Zero + CSV import module. Tagging of the srpELTeC corpus texts was done using the TreeTagger and a linguistic model developed for the Serbian language (Utvić, 2011). Automatic segmentation, tokenization, lemmatization, and Part of Speech (PoS) tagging were performed during the corpus import. Within the TXM environment the results of the tagger are treated as lexical units, namely: `n` (numerical position of the word form in the corpus), `srlemma` (the lemma associated with the token by automatic annotation using the TreeTagger program), `srpos` (PoS associated with the token by automatic annotation using the TreeTagger program) and `word` (concrete token realization in the text) (Example 1).

Example 1. *... из нашега* ***друштвеног*** *живота ...* 'from our **social** life' (part of the text from Lazar Komarčić's novel *Jedna ugašena zvezda*)
`n`: 52.726
`srlemma`: друштвен
`srpos`: A
`word`: друштвеног

The import module generates an XML-TXM format based on the XML-TEI text representation, used as the basic model to represent corpus annotations in the TXM. In addition to building the XML-TXM format, the conversion and generation of the CWB format, used for corpus search applying CQP queries, were also performed. The corpus analysis methodology that the TXM environment enables will be described in the next section.

2.2 Textometric methods in the TXM environment

The frequency is the essential parameter within a corpus, indicating how many times a lexical unit appears in a particular corpus context (Dobrić, 2009). Frequency data serves as the basis for conducting various statistical analyzes, providing an empirical foundation for deriving theories about a language phenomenon.

The primary corpus method is the production of frequency lists. Comparison of absolute frequencies of lexical units (exact number of occurrences in the corpus) can be useful and gives an initial impression of the contrast that exists among the corpus parts, but for the comparison of frequencies in parts of different sizes, it is necessary to normalize, or express the frequencies by a common factor – relative frequency. It would be expected that the relative frequency is calculated as the ratio of the absolute frequency of

the lexical unit and the total number of units in the part of the corpus. The calculated mean value is a mathematical expectation for the normal (Gaussian) distribution of probability. However, the appearance of lexical units in some parts of the corpus is not necessarily consistent with the normal distribution. Pierre Lafon (Lafon, 1984) noticed that the probability of occurrence of lexical units is consistent with the hypergeometric (negative binomial) distribution. The probability that lexical unit A, which is part of corpus vocabulary V, will occur f times in corpus part p of length t, taking into account the total number of occurrences of this unit F in the whole corpus of length T, proposed in (Lafon, 1980), is calculated by the formula:

$$Prob_{specif}(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}, \text{ where}$$
$$C_n^k = \frac{n!}{k!(n-k)!}$$
$$n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$$

The calculation of the *specificity score* based on the hypergeometric distribution in the TXM environment shows the probability of a lexical unit occurring in a particular part of the corpus. The TXM also provides a graphical 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. Thus, it is possible to identify significantly common (positive keywords) or significantly rare (negative keywords) occurrences of a lexical unit in parts compared to the whole corpus, which is a useful starting point for making assumptions about text keywords, domain, or authorship.

Besides the specific frequencies, another standard textometry method is a correspondence factor analysis (Benzécri, 1973). The principles of correspondence factor analysis, developed within the French school „Analyse des Données", have been used to analyze the corpus, but also many other data types (Beaudouin, 2016). This statistical technique enables the display and review of the data set, that is, the interdependencies that exist between corpus and lexical units, in a two-dimensional graphical form.

The initial idea was to discover the patterns of interrelations between two sets of elements recorded in a tabular form. For the sample of textual corpora, if the table columns and raws contain texts and words respectively, at the intersection of columns and rows there are indicators of the presence or word frequency in the text (like word frequency, the specific word frequency). The information contained in the matrices can be synthesized using the

data analysis algorithms. Factor analysis aims to re-organize the matrices containing the maximum amount of information. In other words, the basic idea of conducting correspondence factor analysis is to simplify a complex data set (data cloud) and to find ways to present as much information as possible in a smaller space. Firstly, the gravity center and dispersion of the cloud is calculated. The factor planes and the main axes of the dispersion are constructed in the following step. The points are projected on these planes, and their coordinates on these axes are factors. The plane defined by the first two axes can produce the best cloud projection, which minimizes the loss of information. The main goal is to visualize the distance between the attributes, that is, from the random distribution.

Correspondence factor analysis is often combined with cluster analysis, i.e. hierarchical classification based on the coordinates of the factor axes elements. This classification method serves to identify homogeneous subgroups of texts and words. Cluster analysis, applied along with factor analysis, enables a better understanding of the data and simplifies its interpretation.

3 The results of the textometric analysis

3.1 Corpus general information and frequency

The observed srpELTeC corpus includes texts containing a total of 935.902 words. Regarding the representation of the texts of a particular author, the most extensive parts of the corpus are the texts of Milutin Uskoković, Lazar Komarčić, and Jelena Dimitrijević, while the part containing Milica Janković's novel includes only 8.279 words (Figure 1). The results show that in this corpus, which has 78.542 tokens, 32.604 lemmas were used. The most frequent 20 tokens, carrying little semantic information, account for almost 30% of the total number of concrete realizations in the corpus, while 42.004 tokens appear only once in the corpus (hapax), making slightly more than 50% of the total number of tokens. Table 2 shows the different words of the srpELTeC corpus (`word`), their exact number of occurrences in the corpus (absolute frequency F) and rank in the frequency list sorted by descending frequency (rank). The registered most common words, as well as in the case of the Corpus of Contemporary Serbian Language (SrpKor) (Utvić, 2014), are functional words from closed classes of words such as prepositions, conjunctions, auxiliary verbs, or pronouns.

A frequency list of specific word types is generated based on PoS tags. Table 3 shows the values of the `srpos` attributes, their absolute frequency

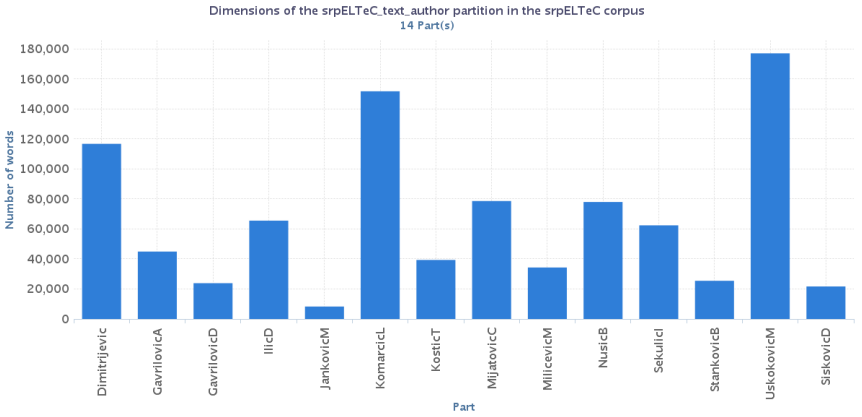


Figure 1. Dimensions of the srpELTeC corpus parts created based on authorship

.

(F) and rank in the frequency list sorted by descending frequency (rank). Aside from nouns and verbs that dominate the texts of the srpELTeC corpus, pronouns also emerge as a morphological category, whose complexity of use and expressive possibilities would be interesting to explore with the TXM tool. Considering the fact that the use of the personal pronoun is arbitrary in the Serbian language because the given verb form also indicates the person, which characterizes the neutral expression style, the high frequency of the pronouns is inherent in stylistically specific literary texts, such as the texts of the srpELTeC corpus (Katnić-Bakaršić, 1999).

Based on frequency data, the TXM allows visual representation of the frequency of particular linguistic phenomena throughout the corpus or parts thereof created based on the existing structural units. Hence, it is easy to notice parts of the corpus and frequency of particular words or expressions. An illustrative representation of the frequency and position of using the lemmas *васељена* 'universe', *црква* 'church', *љубав* 'love' and *девојка* 'girl' in the texts of the entire srpELTeC corpus is given in Figure 2. For example, the lemma *љубав* 'love' is most often mentioned in the novels *Babadevojka*, *Došljaci* and *Đakon Bogorodičine crkve*, in which love is one of the dominant motifs, while the significant use of the lemma *васељена* 'universe' is observed exclusively in the first Serbian science-fiction novel *Jedna ugašena zvezda* of Lazar Komarčić. The lemma *црква* 'church' is mostly used by Nušić at

srpELTeC			SrpKor		
rank	word	F	rank	word	F
1	и	28.545	1	и	4.330.865
2	је	25.422	2	је	4.103.542
3	се	21.128	3	у	3.513.009
4	да	18.932	4	да	3.261.285
5	у	14.932	5	се	2.107.336
6	на	9.233	6	на	1.751.270
7	не	6.721	7	за	1.381.402
8	а	5.642	8	су	1.258.361
9	од	5.234	9	од	919.922
10	што	5.011	10	са	779.469
11	су	4.935	11	а	740.476
12	као	4.857	12	који	650.144
13	за	4.460	13	не	612.218
14	па	4.253	14	о	517.105
15	то	4.132	15	ће	505.643

Table 2. The first 15 rows of the srpELTeC and SrpKor corpora frequency lists

the beginning of his work *Opštinsko dete*, unlike Isidora Sekulić's Đakon Bogorodičine crkve, where this lemma is evenly mentioned throughout the entire novel.

3.2 Specificity score

The frequency of nouns, adjectives, verbs, and adverbs in the Serbian ELTeC corpus is shown in Table 3. Their frequency in the texts of a particular author (f), along with the *specificity score* (S) of the given word type for the entire corpus, is shown in Table 4. The frequency distribution of these types of words in the srpELTeC corpus, partitioned based on authorship, is illustrated in Figure 3. The results reveal that the adjectives are extremely specific for the texts of Lazar Komarčić (S_A=205, 6), whereas Tadija Kostić and Andra Gavrilović use nouns much more often than other authors whose works are included in this corpus (S_N=162, 7 and S_N=83, 6, respectively). On the other hand, the verbs are far less used by Lazar Komarčić compared to their degree of use throughout the corpus, which is presented by the high negative value of the *specificity score* (S_V=−104, 4). The particularly

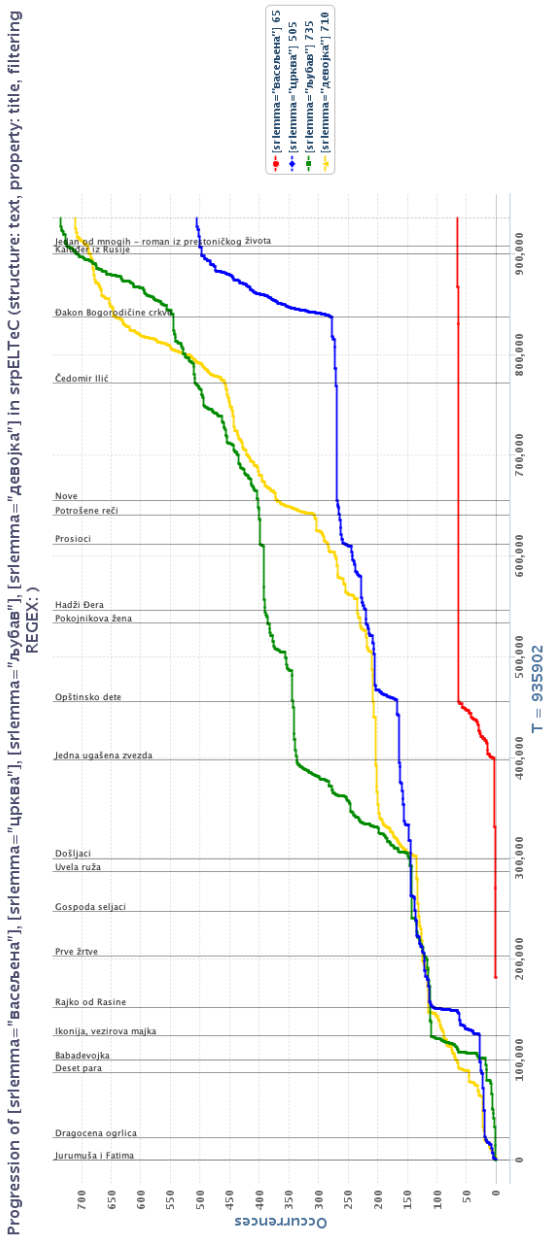


Figure 2. Graphic representation of the specific lemmas' use in the srpELTeC corpus texts

rank	srpos	F
1	N (noun)	192.865
2	V (verb)	173.994
3	PUNCT (punctuation)	103.824
4	PRO (pronoun)	97.239
5	CONJ (conjunction)	90.248
6	A (adjective)	64.242
7	PREP (preposition)	62.584
8	ADV (adverb)	50.628
9	SENT (sentence end marker)	49.184
10	PAR (particle)	36.307
11	NUM (number)	9.208
12	UNDEF (undefined)	2.272
13	? (non-Serbian words or suffixes in compounds)	2.033
14	INT (interjection)	680
15	ABB (abbreviation)	527
16	RN (Roman numeral)	46
17	PREF (prefix)	21

Table 3. Frequency list of position attribute `srpos` possible values in the srpELTeC corpus

low frequency of adjectives is observed in the novels of Branislav Nušić and Dragutin Ilić (S_A=−120,5 and S_A=−82,4), while the nouns, given their degree of use in other parts of the corpus, are far less represented in the stories of Borisav Stanković (S_N=−108).

3.3 Correspondence factor analysis and cluster analysis

In order to simplify the presentation and provide better visibility of the obtained correspondence factor analysis results, the srpELTeC corpus is divided into four parts only (which is the minimal number of corpus parts over which a correspondence factor analysis can be carried out), based on the gender of the author and the century in which his work was published. Therefore, the following corpus parts are marked: *f19* – works of the female authors, published in the 19$^{\text{th}}$ century; *f20* – works by the female authors, published in the 20$^{\text{th}}$ century; *m19* – works of the male authors, published in the 19$^{\text{th}}$ century; and *m20* – works of the male authors, published in the

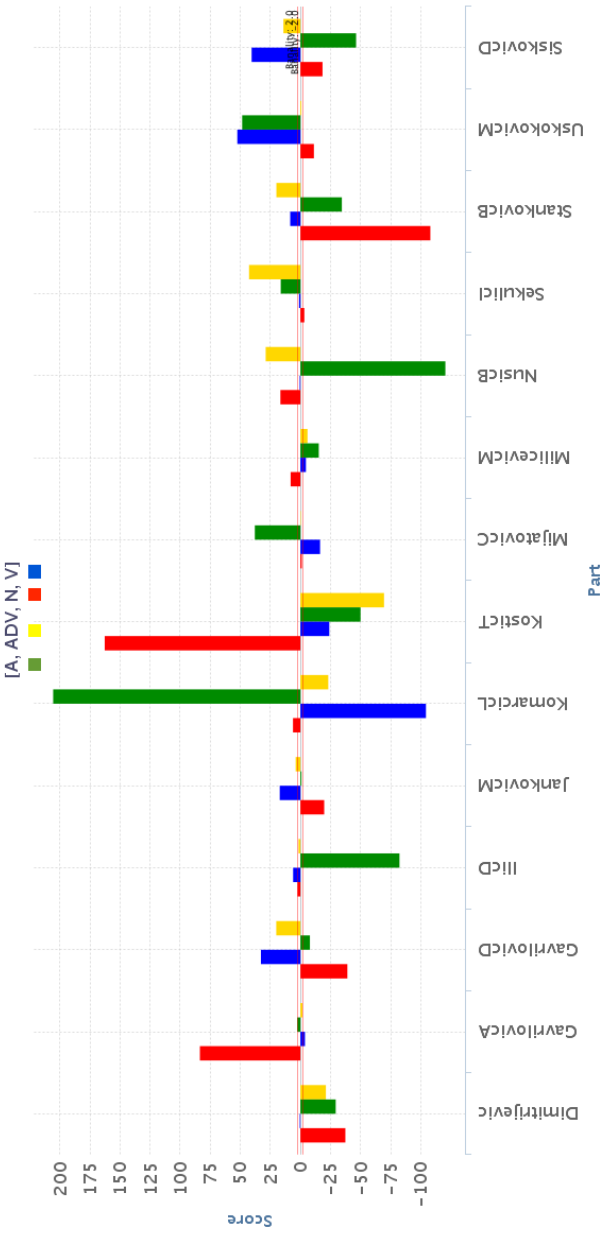


Figure 3. The specificity of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) use in the srpELTeC corpus by authors

Author	f_N	S_N	f_V	S_V	f_A	S_A	f_{ADV}	S_{ADV}
Uskokovic M	35331	-11.2	35446	52.5	13502	48.4	9577	-0.8
Komarcic L	31869	6.2	25443	-104.4	13193	205.6	7481	-23.1
Dimitrijevic J	22330	-37.4	21938	0.5	7066	-29.3	5687	-21.2
Nusic B	16926	16.6	14675	0.6	3812	-120.5	4952	28.9
Mijatovic C	15985	-1.2	13862	-16.3	6259	37.9	4276	-0.4
Ilic D	13728	2.4	12740	6.1	3319	-82.4	3684	1.6
Sekulic I	12497	-3.3	11826	1.1	4772	16.4	4183	42.7
Gavrilovic A	10879	83.6	8125	-3.8	3215	2.7	2356	-1.7
Kostic T	10276	162.7	6606	-24.0	1983	-50.0	1411	-69.5
Milicevic M	7464	8.1	6140	-4.6	1981	-15.3	1680	-5.9
Gavrilovic D	4105	-39.0	5197	32.8	1417	-7.9	1635	20.1
Stankovic B	3867	-108.0	5126	8.5	1268	-34.4	1731	20.0
Siskovic D	3937	-18.4	4838	40.6	981	-46.3	1445	14.1
Jankovic M	1371	-19.8	1860	17.2	539	-0.9	530	4.0

Table 4. The frequency of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) by authors and their *specificity score* for the whole corpus

20$^{\text{th}}$ century. The size of the corpus texts, depending on the gender of the author and the century in which the work was published, is shown in Figure 4, where it can be seen that the part containing the works of the 19$^{\text{th}}$-century female authors is significantly lower than the other parts.

Data on the frequency of nouns, adjectives, verbs, and adverbs in the srpELTeC corpus, divided into parts based on the gender of the author and the century in which the work was published, are shown in Table 5. For each word type, the total number of occurrences in the whole corpus (F), the total number of occurrences in the texts of a particular part (f) and the *specificity score* (S) of the given word type with respect to the entire corpus, are given.

The specific use of word types characteristic for the male or female authors and period (19$^{\text{th}}$ or 20$^{\text{th}}$ century) is also presented in Figure 5. In works published by the 19$^{\text{th}}$-century male authors (part *m19*) nouns are far more prominent than in other parts (S_{m19}=129, 4128), while the use of the verbs is more specific for the part *f19* (S_{f19}=32, 8464), consisting of Draga Gavrilović's novel published in 1887. The specificity of the verbs and adverbs usage is statistically significantly negative in works published by male

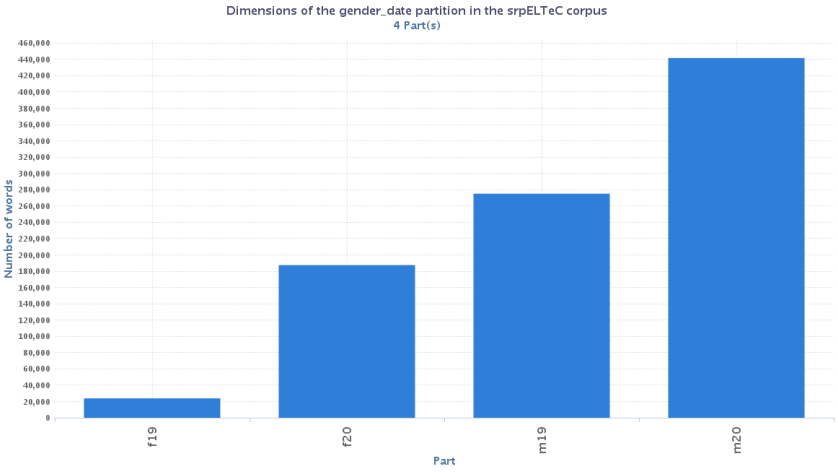


Figure 4. Dimensions of the srpELTeC corpus partitioned based on author's gender and the century in which the work was published

authors in the 19$^{\text{th}}$ century (S_{m19}=−48, 6432 for verbs, S_{m19}=−53, 5126 for adverbs).

Based on the frequency of word types, and according to the χ^2 distribution, correspondence factor analysis was carried out and presented in a two-dimensional graphical form (Figure 6). The obtained factorial map shows that verbs and adverbs are more commonly used in the parts *f19* and *m20*, so they are positioned on the same side of the horizontal axis (the *specificity score* has a positive value). On the opposite side is the part *m19*, which has a markedly negative *specificity score* of the use of verbs and adverbs, but also the part *f20*, whose *specificity score* of the use of verbs and adverbs is positive, but indicates a much lower use of the verbs and adverbs in this part in relation to parts *f19* and *m20*. For this reason, the parts *m19* and *f20*, characterized by a smaller representation of verbs and adverbs, are in opposite quadrants of the vertical axis. Looking at the vertical axis, we can see that there is a part *m19* on one side, in which an extremely high *specificity score* for the nouns is recorded, while the parts in which the nouns are far less represented are located on the opposite side.

The visual representation of the results of the correspondence factor analysis conducted over the corpus divided into author-based parts, and in terms

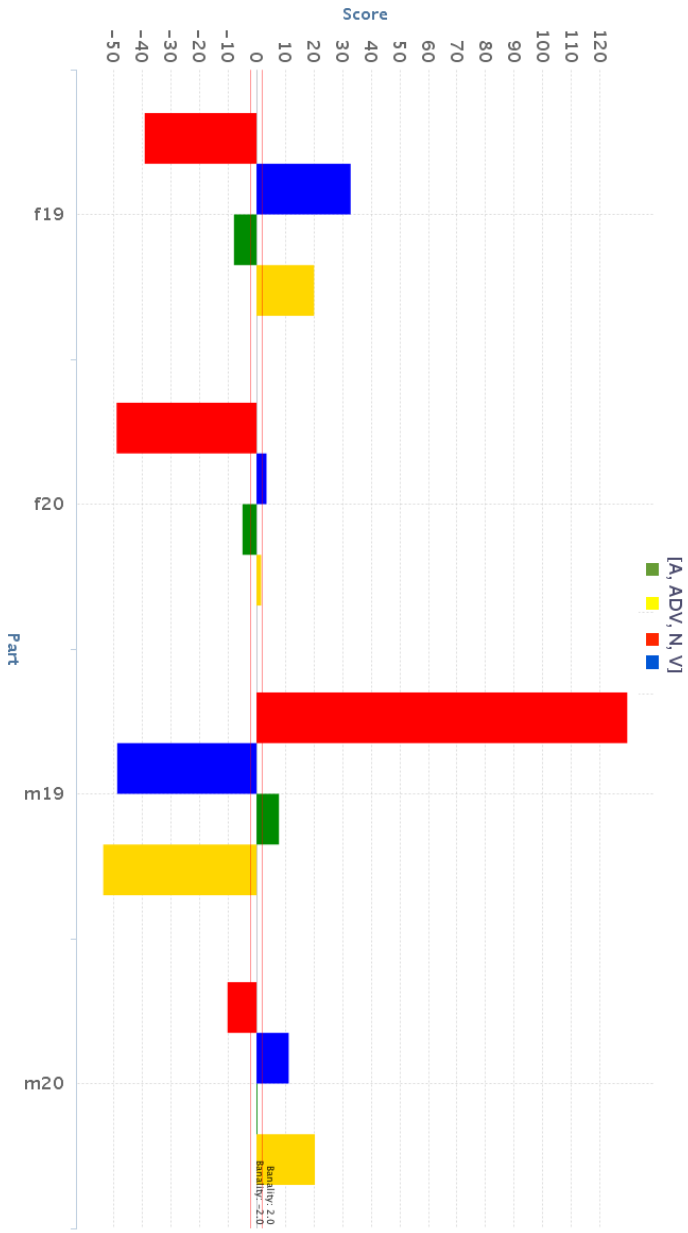


Figure 5. The specificity of certain word types in the srpELTeC corpus by author's gender (m – male or f – female) and the period (19th or 20th century)

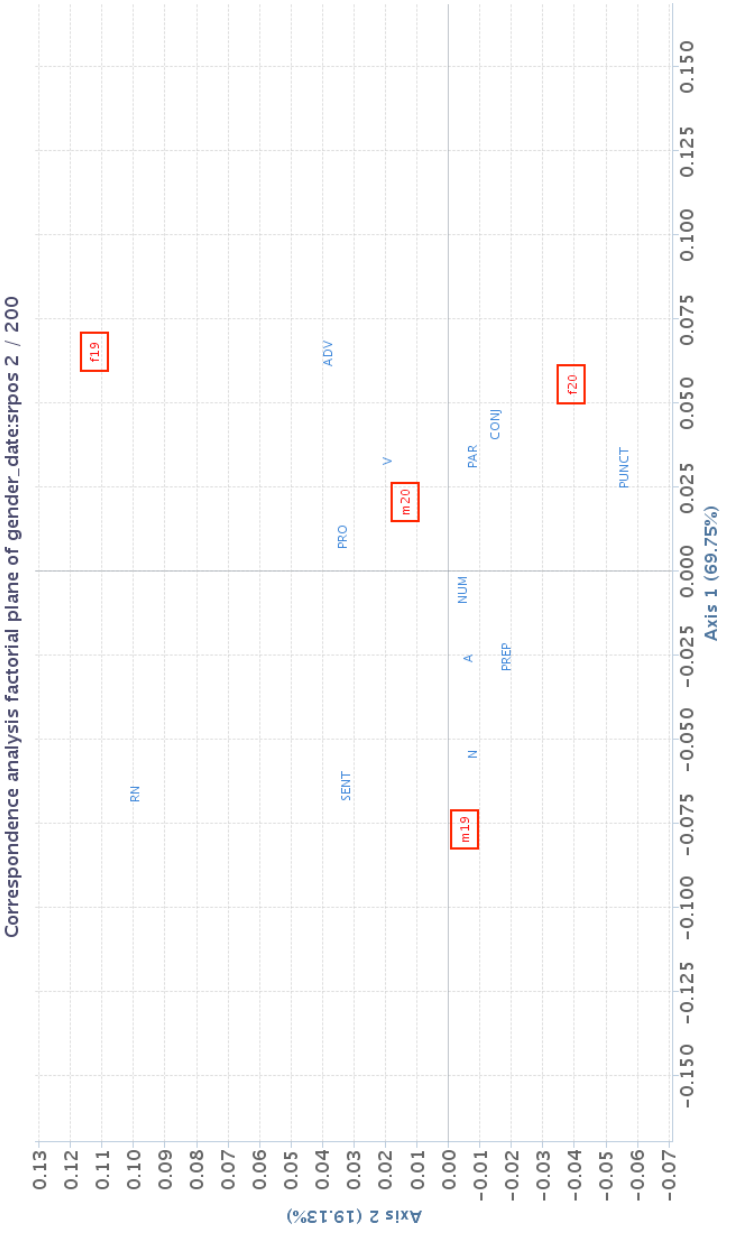


Figure 6. The result of the correspondence factor analysis applied over the corpus partitioned based on author's gender and the publication date

Unit	F	f_{f19}	S_{f19}	f_{f20}	S_{f20}	f_{m19}	S_{m19}	f_{m20}	S_{m20}
N	190565	4105	-39.0398	36198	-48.8455	60820	129.4128	89442	-10.1284
V	173822	5197	32.8464	35624	3.4805	49007	-48.6432	83994	11.2368
A	63307	1417	-7.8845	12377	-4.9007	19383	7.8305	30130	0.3083
ADV	50628	1635	20.0939	10400	1.6149	13477	-53.5126	25116	20.3649

Table 5. The frequency and the *specificity score* of nouns (N), verbs (V), adjectives (A) and adverbs (ADV) by parts created based on author's gender and the period

of the specificity of the used lemmas, is far more complex (Figure 7). The analysis conducted in this way enables the identification and study of laws and trends, otherwise not easily noticeable due to a large amount of diverse data. For example, Milutin Uskoković and Milica Janković use lemmas *живот* 'life', *љубав* 'love', *волети* 'to love', *мисао* 'thought' and *осећати* 'to feel' far more often than other authors, and are placed in the same quadrant of the factorial map. On the opposite side of the horizontal axis, in the upper left quadrant, is the lemma *црква* 'church', as well as the author Isidora Sekulić, who uses it more often than other authors. On the other hand, since Isidora Sekulić, apart from the lemma *црква* 'church', very often uses the lemmas *љубав* 'love' and *живот* 'life', as well as the authors Uskoković and Janković, the mentioned authors and the used lemmas are on the same side of the vertical axis. The results of this analysis are gaining their full significance only after an adequate expert interpretation, which goes beyond the scope of this paper.

In the last step, cluster analysis of the matrix obtained by the previously conducted correspondence factor analysis was also performed. The tree diagram (Figure 8) shows the hierarchical grouping based on the relations existing between the author's texts and the lemmas used in them. This classification of texts provides a better understanding and simpler interpretation of the correspondence factor analysis results.

4 Conclusion

This paper presents the current version of the srpELTeC corpus, consisting of Serbian prose literature from the late 19$^{\text{th}}$ and early 20$^{\text{th}}$ centuries. To illustrate the possibilities of the textometric approach, the analysis of the srpELTeC corpus was performed within the TXM programming environment,

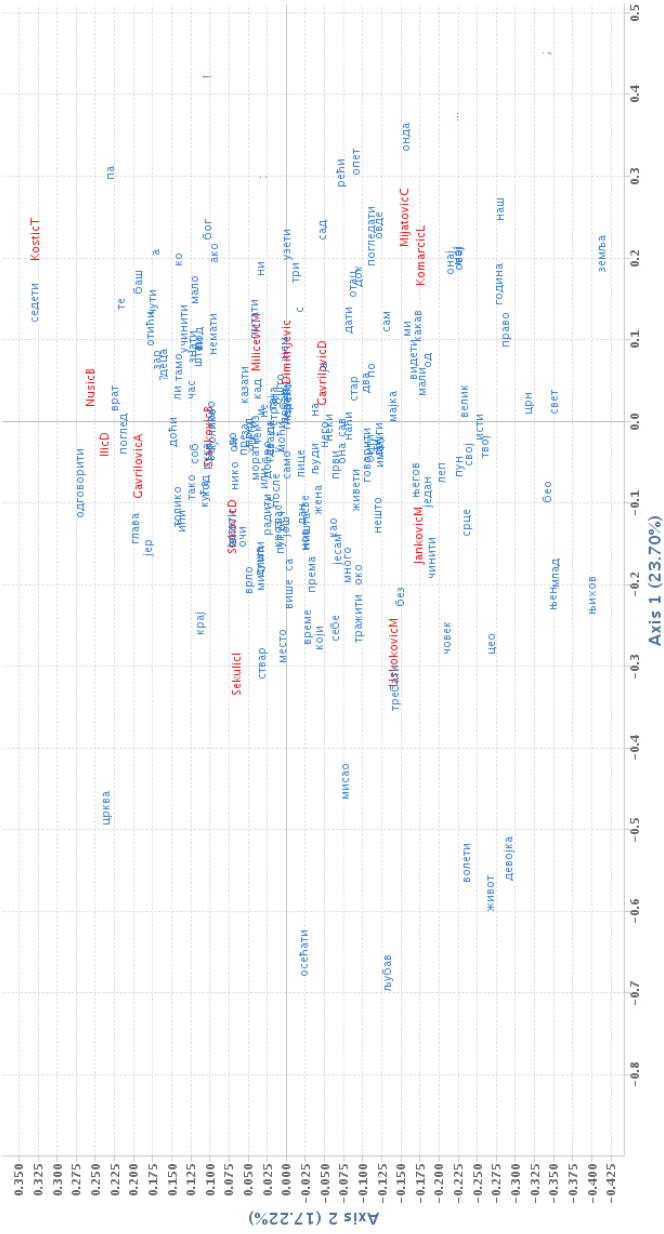


Figure 7. The result of the correspondence factor analysis applied over the corpus author-based parts

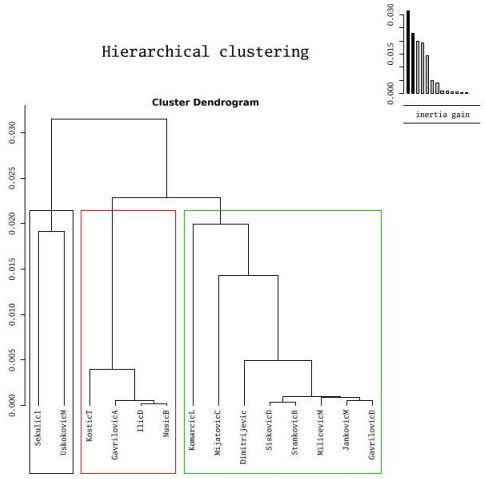


Figure 8. Cluster analysis conducted over the correspondence factor analysis results

presenting the visualization possibilities of the obtained results. The srpELTeC corpus analysis, or some scenarios for the TXM tools application, has no other purpose but to demonstrate the possibilities of using a tool that indicates the significance of textuality and suggests some directions of analysis of those parts that expose and arise from the projections of the corpus itself.

The textometric approach has been used for a long time as a useful method for analyzing corpora of different fields of humanities and social sciences. The laws and conclusions derived from textometric research are based on qualitative and quantitative analysis. The qualitative analysis allows establishing initial hypotheses, which can then be tested on a larger sample by quantitative analysis. The purpose of quantitative or statistical methods is to point out those places in the text that differ and deviate in particular properties. This different way of reading the texts enables asking new questions in the right places, not to give answers, but to identify places that need to be read again and further analyzed, leading to valid interpretation.

Acknowledgment

The author thanks to the COST Action 16204 – *Distant Reading for European Literary History* support, which made possible this research and the author's visit (STSM-CA16204-42562) to the IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) laboratory, École Normale Supérieure de Lyon, France. The author especially thanks to the hosts Serge Heiden and Bénédicte Pincemin for their hospitality and helpful comments and suggestions.

References

Beaudouin, Valérie. "Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis.". *Glottometrics* Vol. 33 (2016): 56–72

Benzécri, Jean-Paul. *L'analyse des données*. Vol. 2, Dunod Paris, 1973

Dobrić, Nikola. "Korpusna lingvistika kao osnovna paradigma istraživanja jezika". *Naučnostručni časopis za jezik, književnost i kulturu Philologia* Vol. 7 (2009): 47–57

Evert, Stefan and Andrew Hardie. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium", 2011

Guiraud, Pierre. *Les caractères statistiques du vocabulaire*. Presses universitaires de France, 1954

Guiraud, Pierre. *Problèmes et méthodes de la statistique linguistique*. D. Reidel Publishing Company, 1959

Heiden, Serge. "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*, 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010

Heiden, Serge, Jean-Philippe Magué and Bénédicte Pincemin. "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*, Vol. 2, 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, 2010

Hunston, Susan. *Corpora in applied linguistics*. Ernst Klett Sprachen, 2002

Katnić-Bakaršić, Marina. *Lingvistička stilistika*. Budimpešta: Open Society Institute, 1999

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

Lafon, Pierre. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots. Les langages du politique* Vol. 1, no. 1 (1980): 127–165

Lafon, Pierre. *Dépouillements et statistiques en lexicométrie*, Vol. 24. Paris: Slatkine-Champion, 1984

Lavrentiev, Alexei, Serge Heiden and Matthieu Decorde. "Analyzing TEI encoded texts with the TXM platform". In *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*, 2013

Lebart, Ludovic and André Salem. *Analyse statistique des données textuelles: questions ouvertes et lexicométrie*. Dunod Paris, 1988

Lebart, Ludovic and André Salem. *Statistique textuelle*. Dunod Paris, 1994

Muller, Charles. *Initiation au méthodes de la statistique linguistique*. Classiques Hachette, 1973

O'Keeffe, Anne and Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010

Pincemin, Bénédicte. "Sept logiciels de textométrie", , 2018, URL https://halshs.archives-ouvertes.fr/halshs-01843695, working paper or preprint

Schmid, H. "TreeTagger-a language independent part-of-speech tagger". *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/* (1994), URL https://ci.nii.ac.jp/naid/20000989946/en/

Utvić, Miloš. "Izgradnja referentnog korpusa savremenog srpskog jezika". PhD. thesis, Univerzitet u Beogradu, Filološki fakultet: Beograd, 2014

Utvić, Miloš. "Annotating the corpus of contemporary Serbian". *INFOtheca* Vol. 12, no. 2 (2011): 39–51

Medical domain document classification via extraction of taxonomy concepts from MeSH ontology

UDC 004.82:025.43MESH
DOI 10.18485/infotheca.2019.19.1.3

ABSTRACT: This paper is a result of a task presented to attendants of *Keyword Search in Big Linked Data* summer school, that was organized by Vienna University of Technology, under the *Keystone COST* action in the summer of 2017. It presents a specific approach to the classification via creation of minimal document surrogates based on the US National medical library's MeSH ontology, which is derived from the Medical Subject Headings thesaurus. In a series of previously classified medically related texts, which are the bases for the task, all of the significant terms are located and replaced with taxonomical references from the MeSH ontology. Extracted references are used for the classification within the ontology using a rather simple algorithm and the results are evaluated in compresence to previous manual classification of the same documents.
KEYWORDS: document classification, MeSH, ontology, information extraction.

PAPER SUBMITTED: 21 April 2019
PAPER ACCEPTED: 30 August 2019

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Belgrade, Serbia

Mauro Dragoni
dragoni@fbk.eu
Fondazione Bruno Kessler
Trento, Italy

1 Introduction

1.1 About the task

This paper describes an attempted solution of an assign given during a one-day hackathon by the lecturers of the summer school Keystone 3^{rd} training

school: *Keyword Search in Big Linked Data*,[1] organized in Vienna from 21–25 August, 2017 under the COST action IC1302 Keyword search in Big Linked Data. The task was to classify 10.000 given documents originating from the digital collection of the US National medical library (Figure 1), whereas the classification to be used was predefined in the MeSH (Medical Subject Headings) ontology (Dragoni, 2017).

> *...Goserelin in the adjuvant treatment of breast cancer An update of the Zoladex Early Breast Cancer Research Association (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). Goserelin was found to be better ... Results were presented by the Austrian Breast and Colorectal Cancer Study Group comparing ...*

Figure 1. A fragment of one of the documents to be classified.

The classification presented in this paper was done according to the medical subjects from the MeSH ontology, version for 2016.[2] Ontology can be queried via web,[3] where you can get predefined queries, the ones for classes and predicates being aong them, or where other data can be obtained with the new SPARQL queries.

The documentation, RDF triplets and the case example download are available online,[4] and there is also an option for previewing the predicate via access point, where the predicates can be seen in tabular form, as well as their descriptions and XML labels, which is especially important if a local copy of the MeSH ontology is used[5] Ontology consists of 56,309 medical concepts, described and systematically classified in a hierarchical tree (Figure 2).

The concepts from the ontology are hierarchically collated, and each has an assigned identifier consisting of blocks of digits, separated by colons, that describe the parent concepts in descending order, from highest to lowest in the hierarchy. In this case, the classification classes relate to the second level

[1] Big Linked Data (on-line)
[2] The classification of documents from a medical domain based on ontologies is the subject of research by a number of teams, using different approaches, but as an ontology, MeSH is most often used.
[3] Access point (on-line)
[4] RDF triplets (on-line)
[5] MeSH ontology (on-line)

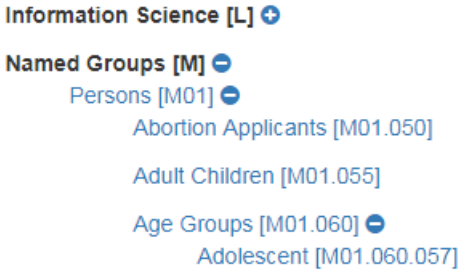


Figure 2. View of a hierarchical tree cut.

of the tree hierarchy – there are a total of 1,718 - and are recognized by two-block identifiers, for example [M01.055] Adult children, where the first block – M01 indicates that it has a parent node [M01] Persons, and the second block 055 is unique among the sibling nodes (Figure 2).

1.2 Classification introduction

The problem of classifying documents in general occurs in two variants: classification in a well-known, restricted domain of classes, and in an unknown one. For both, the problem is reduced to calculating the similarity of documents, usually with use of the so-called Dice[6] index or coefficient.

$$sim(D_i, D_j) = \frac{2|S_i \bigcap S_j|}{|S_i| + |S_J|} \tag{1}$$

Dice's equation tells us that if S_i is a set of terms from document D_i, and S_j is a set of terms from document D_j, then this index can be defined as a double the number of common terms divided with the total number of terms in both documents (if S is a set, $|S|$ is a number of terms in that set). If documents do not have any common terms then $sim(D_i, D_j)$ is equal to 0 which reflects the minimal similarity of the two documents, and if two documents have exactly the same set of terms assigned then $sim(D_i, D_j)$ is equal to 1 which reflects maximum similarity. When a domain of classes is known and limited, the problem is reduced to finding the appropriate class

[6] Lee Raymond Dice – American biologist (1887-1977)

with the largest sim(document, document class) value, i.e. the class most similar to the document that is the subject of the classification.[7]

The problem that arises in calculating the coefficient of similarity between texts is high computer cost, which must be paid either in processing power or high execution time. For this reason, the first step in classifying (and indexing) is most often the creation of a document surrogate. Usually, documents are translated into the word-vector space, or a frequency index. Sometimes words in the index are further derived using stemming or lemmatization, and sometimes by replacing synonyms or hypernyms, in order to further reduce the surrogates and speed up the execution time. For a qualitative classification process, it is necessary to create a surrogate which properly represents the document.

(Trieschnigg et al., 2009) and (Elberrichi et al., 2012) tested a few methods based on MeSH classification based on either MeSH ontology or thesaurus. Created classifiers used:

- MeSH Thesaurus only ('Thesaurus-oriented' classifiers);
- Training set to build explicit models for each MeSH concept ('Concept-oriented' classifiers);
- Manually created document annotations, like ordinary text classifiers, to determine the appropriate concept ('K-Nearest Neighbor' classifier);
- Hybrid and hand-refined systems that combine multiple approaches – 'Hybrid' classifiers.

In both papers, it was concluded that the K-Nearest Neighbor Classifier (KNN) produces the best results, but despite its advantages, it is significantly slower than the thesaurus-based classifiers, and with the growth of a set of test documents, its performance further decreases, which was not desirable in solving our task.

In this paper, we experiment with a simple classification approach to evaluate the importance of timely management of large amounts of data, as well as the usable value of semantics stored in the MeSH ontology. The goal was to create a classifier that would be quick and simple, in order to solve the problem of the large amount of text that needed to be classified. A drastic summarization of documents and the classes themselves was applied. Classes (concepts of the second level of ontology) were reduced to a single term – their name. On the other hand, the documents were reduced only to the occurrences of terms (concept names from MeSH ontology) that, with

[7] IR notes (on-line)

the simple mapping (stored in their identifiers in the ontology itself), are identified with a term that denotes a class, their parent object. This greatly facilitates and speeds up similarity calculations, as each class now has only one term. In this way, the document will always be classified by the Dice index into the class whose (only) term occurs most often in it, thus avoiding a large amount of computation and reducing the task to finding the most frequent term in the surrogate of the text.

2 Experiment setting

The aim of the experiment was to test the possibility and success of classification of medical documents based on taxonomy from the MeSH ontology and a rule-based system managing the appearance of terms related to concepts from the MeSH ontology in the documents to be classified. The course of the experiment can be divided into five follow-up steps.

1. **Extraction of taxonomy from MeSH ontology using SPARQL query**. This stage was necessary in order to snap together the list of identifiers and determine the taxonomic position of the concepts used in the documents, as well as the relative position of their nodes in the hierarchy.
2. **The conversion of documents into vectors of identifiers using concepts occurring within them**. This stage allows the assignment of attributes that are directly and inextricably linked to the classes in which these documents should be classified.
3. **Noise removal**. This stage should enable and provide better results for the document classification.
4. **Document classification based on their identifier vectors and a simple set of rules**.
5. **Evaluation of document classification performance for each of the sets used**. This stage allows us to reflect on and compare different classification rules, as well as to determine whether some rulesets can (and to what extent) be considered successful.

In the following chapters, these experiment stages will be described in more detail, in order to get a better insight into the methods used and the results obtained.

2.1 Extraction of taxonomy of concepts from MeSH ontology

Extracting the matrix of the concept names and their identifiers in the classification tree is done using another SPARQL query. Since in this ontology there are triples consisting of the concept, predicate and object of that predicate, which reflects the position in taxonomy, this part of the task is reduced to the extraction of a subject and object for each of these triplets.

First, it was necessary to find the name of the predicate that reflects the position in the taxonomy in the form **[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***.[8] A simple SPARQL query was used, with one ontology concept (mesh2016: D049916) inputted, and it lists all predicates and objects of the MeSH 2016 ontology triple, whose concept is a part of (Figure 3). The query is illustrated on the concept *mesh2016:D049916*.

```
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?predikat ?objekat
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE  mesh2016:D049916 ?predikat ?objekat
ORDER BY ?class
```

Figure 3. SPARQL query used to provide taxonomy of all concepts and MeSH 2016 ontologies.

Based on the query, a set of results is obtained containing, inter alia, the output from which it is concluded that the required predicate is *meshv:treeNumber* because it contains the syllables describing the hierarchy (Figure 4). The data is used in a subsequent query that aims to derive all the names of the concepts and their *meshv:treeNumber* values.

The concept names are derived from the *rdfs:label*, followed by the *mesh:treeNumber* of the same concept. The returned concepts are sorted by the size of the name from the longest to the shortest, in order for them to be searched in the documents without the risk of longest matches not being recognized due to previous recognition of shorter ones (Figure 5).

The result of this query is a CSV file whose rows contain the name (*rdfs:label*) and the taxonomic reference (*treeNumber*) of each concept (Fig-

[8] This regular expression describes a construction that consists of a mandatory part (capital letter, number, digit) and an optional part (dot, number, number, number) that iterates.

rdf:type; meshv:TopicalDescriptor
rdfs:label; Polyplacophora
meshv:identifier; D049916
meshv:dateEstablished; 2006-01-06
meshv:historyNote; 2006
meshv:publicMeSHNote; 2006
meshv:treeNumber; mesh2016:B01.050.500.644.600

Figure 4. Some of the SPARQL Queries 1 outputs, among which are the desired predicate and object

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?naziv ?treeNumber
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE ?koncept rdfs:label ?naziv .
?koncept meshv:treeNumber ?treeNumber
ORDER BY DESC(STRLEN(?naziv))
```

Figure 5. SPARQL query used for provision of a list of the name and position of all the concepts from the MeSH 2016 ontology.

ure 6). This file will be used in the next step, where the concept names are located in the documents and replaced with the node identifiers from the taxonomic tree. It should be noted that both one-word and multi-word units (e.g. Gram-Negative Bacteria) can be found. However, having in mind the order of applying the replacements (from the longest to the shortest term), there will be no wrongful replacement and recognition of only a part of the term.

2.2 Conversion of documents into concept vectors

This stage consists of two steps. First, in all documents, the previously listed concepts are found and replaced with corresponding identifiers, and then the remaining text is removed in order to transform the documents solely into the list of the identifier vector. No normalization of the document or concepts was done, which is sustainable for English which does not have a rich flexible

```
Ganglia;A08.340
Neurons;A08.675
Malleus;A09.246.397.247.524
Cochlea;A09.246.631.246
Eyelids;A09.371.337
Choroid;A09.371.894.223
Tissues;A10
Chorion;A10.615.284.473
Muscles;A10.690
```

Figure 6. Examples of lines from CSV document containing the names and identifiers of concept nodes.

system, but for a morphologically rich language, such as Serbian, previous lemmatization or other kind of normalization of both resources is necessary.

Finding and replacing ontology concepts in the text. As we wanted to find something in the documents (names of the concepts) and then replace it with something else (corresponding taxonomic identifiers), having those two things listed together in a previously generated file, it was possible to directly transform the list from the file directly to C# function that would do it.

This is achieved by through transformation of the CSV file. Character **;** was replaced with string **", "** and strings **doc = doc.Replace("** and **");** were pasted onto the beginning and the end of each row respectively (Figure 7).[9]

For replacement in all the documents to be classified, a second C# code has been prepared. It loads the classification documents one at a time and applies the script generated in the previous step so that the concepts are found by name and replaced by an ontology node identifier. This stage is the longest and the most time consuming because our experiment involves the application of 56,309 term replacements over 10,000 documents, giving a total of 563,090,000 transformations. Further research can go towards using

[9] In hindsight, a similar approach can also generate a different type of substitution based on loops or regular expressions, which would speed up replacements and reduce the effects of multiple parsings of the same document.

```
doc = doc.Replace("Ganglia", "A08.340");
doc = doc.Replace("Neurons", "A08.675");
doc = doc.Replace("Malleus", "A09.246.397.247.524");
doc = doc.Replace("Cochlea", "A09.246.631.246");
doc = doc.Replace("Eyelids", "A09.371.337");
doc = doc.Replace("Choroid", "A09.371.894.223");
doc = doc.Replace("Tissues", "A10");
doc = doc.Replace("Chorion", "A10.615.284.473");
doc = doc.Replace("Muscles", "A10.690");
```

Figure 7. The section of a find and replace script based on the previously generated CSV file (Figure 6).

finite state machines and transducers to solve this problem, which is more complex to implement but performs faster in processing this type.

Transformation of documents into concept vectors (surrogate creation) After the documents were successfully annotated with the identifiers of concepts that appeared in them, it was necessary to clear the documents from the rest of the unpaired text. To prevent this from working individually for each document, they are merged into one, ∼230MB, in size, with new lines as the border between the documents. Information of importance – document names (**[0-9]+[.]txt**), identifiers in them (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***) and tags for new row (**[\r\n]+**) - are found using regular expression (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])*)|([0-9]+[.]txt)|([\r\n]+**), док се све остало уклања.while everything else is removed. This reduced the file size over 450 times (new size: ∼0.5MB).

Upon completion of the transformation, a new file is formed. In it, each new line represents a new document: it begins with the title of the document (without extension) followed by a semicolon, and all the concept identifiers found in it separated by commas (Figure 8).

We will illustrate the transformation of one of the starting documents by steps on a simple example. In a document fragment from 1, some terms were identified (Figure 9), and then replaced by identifiers (Figure 10). It is noted that in the *Colorectal*, part of the word was recognized, and a string **Color** was mistakenly replaced by **G01.590.540.199**.[10] This happened because

[10] Such an error could have been avoided by previous tokenization of the text.

2875592;M01.060.116
2875593;D13.444.308,D13.444.308
2875594;C04.557.465.625.650.510,D13.444.735,D13.444.735
2875595;A01.236,A01.236;A01.236
2875596;D13.444.735
2875598;
2875599;D02.455.612

Figure 8. A fragment of a file that contains the names of documents and identifiers in them.

neither the terms *colorectal cancer* nor *colorectal* are found as terms in the ontology version used. Figure 11 shows the final surrogate of the text from Figure 1.

... **Goserelin** in the adjuvant treatment of breast cancer An update of the Zoladex Early **Breast** Cancer **Research Association** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **Goserelin** was found to be better ... Results were presented by the **Austrian Breast** and **Color**ectal Cancer Study Group comparing ...

Figure 9. A fragment of the original document with the concepts found in MeSH ontology marked.

... **D06.472.699.327.740.320.340** in the adjuvant treatment of breast cancer An update of the Zoladex Early **A01.236** Cancer **H01.770.644 F02.463.425.069** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **D06.472.699.327.740.320.340** was found to be better... Results were presented by the **Z01.542.088 A01.236** and **G01.590.540.199**ectal Cancer Study Group comparing ...

Figure 10. A fragment of the original document with the concepts found in MeSH ontology replaced with identifiers.

...D06.472.699.327.740.320.340;A01.236;H01.770.644;F02.463.425.
069;D06.472.699.327.740.320.340;A01.236;G01.590.540.199;...

Figure 11. Final surrogate of a document fragment.

2.3 Noise removal

Before moving onto classification, it was necessary to detect possible noise in the form of identifier attribution errors, highly frequent concepts or the ones that appear in an excessive number of documents, which are therefore not discriminatory. For each class, a total number of repetitions was calculated, resulting in an uneven distribution (Figure 12).

The first identifiers added to the list of stop words and removed are those relating to geographic locations (listed in ontology under class Z01 - geographic locations as well as homonyms like the term back, which appears in documents 11,440 times, apparently not always to denote parts of the human body. Figure 13, however, shows the uneven distribution of frequency classes even after this removal.

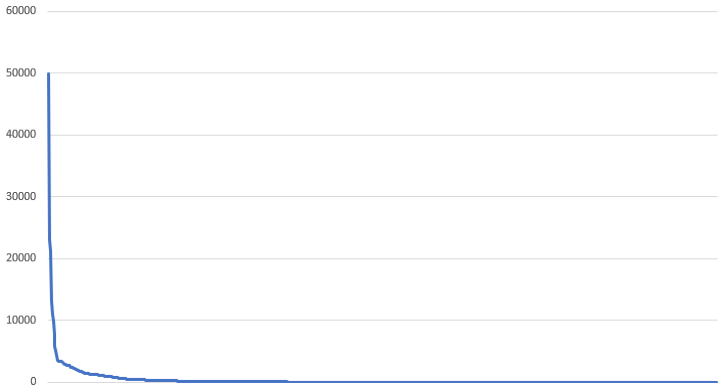


Figure 12. Frequencies distribution before noise removal.

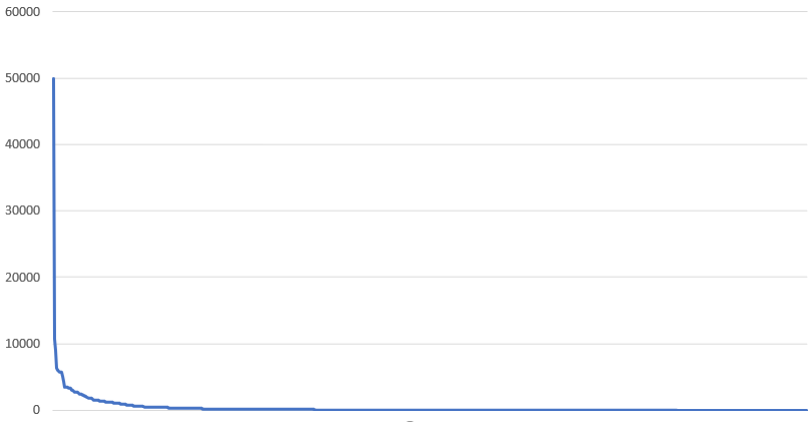


Figure 13. Frequencies distribution after noise removal.

2.4 Classification of documents using identifiers

Two processing procedures were applied to the documents subject to classification, resulting in two test sets. The control method was to replace the identifiers with their class, a more general hierarchical designation. The experimental method also considered the length of the identifiers, so they were replaced by a certain number of iterations of the parent class depending on their length, to test the assumption that the use of more specialized terms was more important for determining the class of documents. So, instead of reducing the identifiers to the first two blocks of digits, their length was taken into account or the depth of each of the concepts in the tree. For example: identifier **D04.345.295.750.650.700** has been replaced using an appropriate regular expression with **D04.345, D04.345**, which is equivalent to the appearance of two concepts belonging to class **D04.345** in that document. The way to map the length of concepts into the corresponding number of repetitions is given in Table 1.The table shows that terms are identified with a maximum of four repetitions (if they have more than eight blocks) of a term denoting a class. After applying these steps, only class identifiers now appear in document surrogates, which should be easily counted.

After the test sets have been successfully created, a simple program is prepared for document classification, which requires a file with inputs indicating the classes (the first two blocks of digits) that are recognized as

number of surplus blocks (over 2)	0	1	2	3	4	5	6	7	8	9	10
number of resulting class iterations	1	1	2	2	2	3	3	3	4	4	4

Table 1. Mapping the number of surplus syllables and the number of resulting class identifiers.

input. The program simply counts the classes that occur in the surrogate of the document and returns the one that occurs most frequently. If there are multiple classes that occur in the document with the same frequency, the class that first appears returns, which is logical because the order of occurrence of the terms is retained in the surrogates. Also, it is necessary for each document to be assigned an identifier, so if a document does not have an identifier assigned, it is assigned one of the most general – **H02.403** – which designates medicine.

When a sequence of identifiers in each document is reduced to one class, it is taken as a result of the classification and forwarded for evaluation.

3 Evaluation and results comparison

Test set	Taking identifier length into account	Precision @10	Average mean Precision	Recall	F-measure
1	yes	0.58	0.0060	0.0696	0.0108
2	no	0.46	0.0057	0.0648	0.0103

Table 2. Test sets used for classification and their results.

Experiments were performed on documents from the TREC Clinical Decision Support 2016 set.[11] The aim was to classify documents based on terms that denote concepts in the MESH ontology and that appear in those documents. The annotation used in the evaluation was manually conducted by an expert team. The usual metrics of average mean precision, recal[12] and F-measure

[11] TREC Clinical Decision Support 2016 (on-line)

[12] Relevant documents not included in the ranking were taken as false negatives.

were applied, as well as the precision@10 metric, which reflects the success of returning relevant documents in the first 10 results of a query.

Table 2 shows us that taking the length of the identifier into account yielded a slight improvement in results (5% improved precision and F-measure, 7% improved recall and 12% improved precision@10).

Considering the values obtained, it can be immediately noticed that the precision is unusually low relative to recall. By applying a more detailed analysis of the data, a very large number of false positives was observed, thus explaining the decreased precision of the given strategy. This result is not surprising, as it concurs with current standards in the field of classification of medical records (Calí et al., 2017). A major problem with concept-oriented information retrieval in the biomedical sphere is the large number of miss-classified documents, leading to a very low response rate. Low precision is thus acceptable in this paper because it is offset by a higher response rate and many relevant documents are returned in the highest positions, with precision@10 values, as high as 58%. Still, there is room for progress here.

4 Conclusion

In this paper we presented an approach to document classification which is based on the creation of the minimal surrogates of those documents. Within medical documents, specific terms are located and replaced with taxonomical references. Extracted references are used for classification using MeSH ontology and a simple algorithm and evaluated against a team of experts.

Preliminary results demonstrated the suitability of the proposed approach within a very complex task. Future work will focus on the decrease of false positive results in order to boost the overall performance of the system.

The classification based on ontologies does not depend on the domain in which it is applied, but it certainly depends on the resources available, specifically the ontology or taxonomy used for the classification (Rakesh et al., 2001).Once established, the system may find wider application. When it comes to the classification of (medical) documents for the Serbian language, it is necessary to prepare resources first. In this regard the International Classification of Diseases in Serbian - *MKB 10* (*Međunarodna klasifikacija bolesti*) (Kolonja et al., 2016) could certainly be of use, where a number of terms is associated with English and Latin equivalents, allowing for the extension of the search for concept names and their retrieval in documents. However, rich Serbian language morphology should be taken into account

and preparation of additional lexical resources specific to the field of medicine would be required in order to normalize text before classification or indexing, which would help to identify more taxonomic terms in documents (Stanković et al., 2015).

5 Acknowledgements

This work was created within the Keyword Search in Big Linked Data summer school, organized as part of the Keystone COST action from 21 to 25 August 2017 at the University of Technology in Vienna.

References

Rakesh et al., Agrawal. "Multilevel taxonomy based on features derived from training documents classification using fisher values as discrimination values", U.S. Patent No. 6,233,575, 2001

Calí, Andrea, Dorian Gorgan and Martin Ugarte. *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, Vol. 10151, 2017

Dragoni, Mauro. "3rd KEYSTONE Summer School", 2017, URL http://ifs.tuwien.ac.at/keystone.school/slides/Dragoni_SemanticSearch.pptx

Elberrichi, Zakaria, Malika Taibi and Amel Belaggoun. "Multilingual Medical Documents Classification Based on MesH Domain Ontology". *International Journal of Computer Science Issues* Vol. 9 (2012)

Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović and Aleksandar Cvjetić. "Development of terminological resources for expert knowledge: a case study in mining". *Knowledge Management Research & Practice* Vol. 14, no. 4 (2016): 445–456

Stanković, Ranka, Cvetana Krstev, Ivan Obradović and Olivera Kitanović. "Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian". In *Semantic Keyword-based Search on Structured Data Sources*, Cardoso, Jorge, Francesco Guerra, Geert-Jan Houben, Alexandre Miguel Pinto and Yannis Velegrakis, 167–181. Cham: Springer International Publishing, 2015

Trieschnigg, Dolf, Piotr Pezik, Viv Lee, Franciska de Jong, Wessel Kraaij et al.. "MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval". *Bioinformatics (Oxford, England)* Vol. 25 (2009): 1412–8

The Implementation of Dais Repository in ISL SASA

UDC 061.6: 811.163.41]:004.738.5
DOI 10.18485/infotheca.2019.19.1.4

Vladimir Živanović
vladimirludwig@yahoo.com
Institute for Serbian Language of SASA
Belgrade, Serbia

ABSTRACT: The paper presents the involvement of the Institute for the Serbian Language of SASA (ISL SASA) into the world of digital repositories. The ISL SASA was one of the first institutions in Serbia to have an institutional repository (2017). Before that, scarce research outputs of the ISL SASA were available online whereas the available resources remained invisible to major international infrastructures. In developing the repository, the role of the ISL SASA librarian has been crucial. It involved repository management, verifying the metadata accuracy, preserving digital content and further dissemination. More than 100,000 pages (2500 full-text records in the repository) were digitized in one year. This has significantly improved the international visibility of Serbian linguistic humanities.

KEYWORDS: Digital Repository, Institutional Repository, DAIS, DSpace, Open Science, Self-Archiving, Digital Humanistics.

PAPER SUBMITTED: 10 May 2019
PAPER ACCEPTED: 18 June 2019

1 Introduction to the digitization process in ISL SASA

The Ministry of Education, Science and Technological Development adopted the Open Science Platform in 2018 (Platforma, 2018), This Platform, which is based on the principles of Open Science and European Commission guidelines in this field, has introduced an obligation for individuals and institutions to make their own scientific production, which is publicly funded, also

publicly available and to deposit it in a digital repository. In order to meet the "transparency of scientific communication and methodology" and to ensure the availability of scientific production, it was necessary to develop the appropriate digital infrastructure, which in Serbia did not exist.

The open approach in this document is defined as "the right of every Internet user to read, download, store, print and use digital content of publications without financial expense". In addition to these rights, users will respect copyrights, i.e. a license that is associated with deposited work and with proper source guidance participate in the dissemination of open science, using some of the "institutional / thematic / national repositories". The type of repository for the deposit of scientific production is not strictly determined, thus leaving it to institutions to find an adequate solution in accordance with their capabilities. Under this Platform, each of the institutions involved is obliged to adopt a "local" policy, that is, an appropriate document on the implementation of the Platform that will regulate this process at the institutional level.

At the same time, during the preparation of this Platform, the Serbian Academy of Sciences and Arts (SASA), in cooperation with the University of Belgrade Computer Center (UBCC), has worked on developing its own repository designed for the Academy and related institutes. The repository was named Digital Archive of the SASA and the Institute – DAIS.[1] According to Branin's definition, repositories represent "models of systems and services designed to collect, organize, store, share and store digital information and knowledge of the institution" (Branin, 2004-2005, 237). The purpose of the DAIS digital repository is to enable the Academy, as well as the Institutes founded by the Academy, to permanently preserve their scientific production, thus publicly presenting the results of its scientific work.

The development of communication processes leads to the increasing availability of scientific papers in full text. General availability is still an ideal to strive for, but the scientific institutions that have been instigated by numerous open source initiatives, as well as the practice of large tech company services (Google Books, etc.) have already begun with setting up their digital repositories. One of the results of the successfully managed repository is the increase of visibility of scientific papers, especially in the field of Serbian humanities. Due to the commercial potential of technical and natural sciences, their visibility in the world greatly exceeds that of humanities. Such

[1] DAIS (on-line)

a disparity is not justified, given that it represents a culture at the global level.

Unlike a standard library, the formation of a digital repository requires technical preparation. DAIS was created by customizing open source software DSpace.[2] Customizing this software implied its localization, adaptation of the user interface, alignment with the guidelines of the OpenAIRE consortium for digital repositories, integration with ORCID and Altmetric.com, and the development of additional external applications that allow normative base, as well as the downloading of records and the massive input and correction of metadata.[3]

The reason for creating such software is based on the development of library management. Baudoin and Branschofsky from MIT argue that university communities depend on their libraries that allow them constant access to research and scientific work, and enable them to seek solutions to the problem of storage and take-over of intellectual work in the long run (Baudoin and Branschofsky, 2004, 32). The implementation and development of software, as well as its adaptation to the specific needs of Serbian scientific production, was performed by UBCC. The computer center is constantly working on maintaining the repository and its continuous improvement.[4]

One needs to note the difference between a digital collection and digital repositories. While a collection remains local – for example, the collection of digital objects that institutions or individuals upload on their own internet sites – it does not meet the standards of interoperability (DOI number, other persistant identifiers, metadata and attributes, structures and standards for describing a document, metadata exchange protocol, etc. (Van de Sompel and Nelson, 2015)) The Digital Repository, however, has presented a digital collection organized in such a way that it can be archived, preserved and disseminated according to its defined goals and protocols compatible with other databases. Its main purpose is "the presentation of intellectual conduct" and the promotion of accessibility (Winter and Bowen-Chang, 2010, 320).

[2] Dspace open source software was developed by the Massachusetts Institute of Technology (MIT) together with Hewlett-Packard in 2002. Its quality is seen in an easy and open approach to various types of digital objects: text, images, audio or video material.

[3] For more on DSpace software, as the backbone for the development of digital repositories, see (Rajović and Ševkušić, 2018)

[4] Customizing open source software to local needs is the biggest challenge in the initial stages of organizing the repository and requires high-quality technical support. The software code is continuously adapting.

The ISL SASA experience in organizing and managing a digital repository can be useful to other institutions involved in organizing digital full-text collections in open access.

2 Structure

In order for an institutional repository to become functional, the filling of repository is the first major obstacle to overcome. That's why in ISL SASA we had two phases of filling. A properly organized repository increases the visibility of the scientific production of a particular institution, thus enabling better citation of researchers (Piwowar and Haustein, 2018).

The first phase is the collection, digitization and installation of the digital content of the journal and the ISL SASA monograph. The history of periodicals of the Institute is very long. Two of the four scientific journals published by the Institute have been published for more than a hundred years.[5] Published volumes of these journals have several hundred volumes. In addition to digitazion of the serial publications in the first phase in the institutional repository, we have digitazed the scientific papers of prominent scientists from the Institute. There are colleagues who are retired and some of them are deceased. By introducing their digitized works, several successful criteria for the functioning of the repository are fulfilled: the good practice of future work is established, it participates in the formation of the customs of future users, and the scientific production is presented in the best light.[6] This also covers the diachrony of the scientific production of the Institute.

The second phase, which represents a synchronous level, concerns the digitazion of the current scientific production, at the latest 18 months from the date of the publication of a scientific article or monograph, and is defined by the Rulebook adopted at the institute level. This fulfills the conditions provided by the Platform of the Ministry.

Institutions that engage in the formation of a repository should define their goals and clearly identify the set of traits of a well-trained staff involved in the process of maintaining a repository (Winter and Bowen-Chang, 2010, 323). Training a personnel to work in the repository can be a challenge.

[5] The first issue of the journal *Srpski dijalektološki zbornik* was published in 1904, *Južnoslovenski filolog* in 1913, the journal *Naš jezik* in 1933, and the *Lingvističke aktuelnosti* 2000.

[6] Unlike in technical sciences, in linguistics, many old scholarly papers do not lose importance because they are still relevant for the research.

Lack of initiative and enthusiasm in accepting a different approach to scientific publications, as well as insufficient awareness of its significance, can be reflected in the first stages of organizing the repository. It is very important that managerial staff, as well as associates of a scientific institution, recognize the importance of establishing a repository.

All researchers in the Institute are obliged to depose their scholarly work in the process of self-archiving. Training is organized for the researchers in workshops that contain practical work (personal input of metadata and deposition of objects), and a detailed explanation is prepared in the PowerPoint presentations that they can use for their own work. Once a researcher deposes work and describes all the associated metadata, the librarian as an administrator checks the accuracy of entries and approves the record, since only he is in charge of subsequently changing metadata or removing digital object.[7]

A metadata collection involves entering basic document data, text summary, and keywords, as well as determining the level of document availability and the type of license. Also, providing data on the project within which the work is funded is mandatory. The same pertains to data on the type and version of the document, the availability of content and licenses, which are entered in accordance with the current guidelines for the digital repositories of the OpenAIRE consortium. The domain that is becoming increasingly important is monitoring copyright compliance and contacting publishers of the journals for copyright regulation.

Furthermore, the librarian uses additional DAIS services, such as a metadata editing service, a normative database, and a service for downloading the already existing digital content from other repositories. The UBCC Development Team has independently developed the Ellena2 application, which includes a normative database, a service for massive metadata correction, and a mass-feed service for metadata. In the early stage of its development, there were two separate applications: Ellena (normative file and massive metadata correction) and MultiLoad (downloading records from other DSpace repositories and massive import of metadata in XML or RIS format). At the beginning of 2019, these two applications were integrated into the Ellena2 application. The reason for the creation of an independent application is the lack of a DSpace platform, which in itself does not contain a normative

[7] The DAIS repository consists of multiple levels of access and content management permissions, and the administrator has the ability to restrict the user's authority when self-archiving, thereby ensuring quality control, accuracy of data and database integrity.

file module. In addition, initial tests have shown that the system for downloading metadata that was already embedded in DSpace does not provide satisfactory results.

Content management allows researchers to use the repository in a wider range than the Open Source Platform requires because DAIS supports the input and processing of multiple types of documents and formats. In addition to scientific articles, chapters in a monograph, monographs, etc., the associates are able to deposit and pre-print versions of their work, doctoral or master thesis, reports and all other contents called "gray" literature.[8] DSpace indexes all readable text from the deposited documents, so it can also be searched. Ranking indexed documents is done through formal parameters, i.e. by relevance, title and date. DSpace supports the storage of virtually all formats. DSpace does not have a system for viewing documents in the way it works in Omeka and other platforms primarily intended for displaying cultural heritage and digitized material. DSpace is designed to present scientific content, whereas stored files can be saved to the user's computer and then opened through appropriate application: PDF, JPG and similar formats are opened in the new browser window. DSpace can be upgraded to display content within the platform itself (for example, listening books), but for users in ISL SASA, it is not necessary at this moment.

3 Practical work

ISL SASA started the work on DAIS towards the end of 2017. Considering the relevance of ISL SASA throughout the Slavic world, a part of the yearbooks has already been scanned within the Google Books service but has not been made publicly available.[9] Several US universities have digitized Serbian journals as part of their collections in cooperation with Google. These digital copies were collected and they represented approximately 50% of the total volume of the Journals. The rest was to be digitized. The work, which involved organizing, scanning, redesigning digital copies, working on the optical character recognition of content, and depositing it in the database, has

[8] For more on scientific "gray" literature see (Ferreras-Fernández and Merlo-Vega, 2015) and (Ćirković, 2018).

[9] One of the major Google Books missions is the organization of world-level information. By digitizing the world's library heritage, Google becomes a universal library of world-class knowledge. For more about the partner program with Google Books, see (Шевкушић, 2013)

begun. This way over 50,000 pages of text were prepared in a year. The lack of infrastructure in certain segments of the work was replaced by cooperation with Google Books: thanks to the fact that ISL SASA made scanned publications available through this service, we managed to reduce the scanned material, which was very large in memory,[10] to the optimal size, and we provided optical text recognition.[11]

In the first phase, we made the entire existing production of the Institute publicly available by depositing every journal on the annual basis, i. e. as a single document, while it remained to deposit each article separately, thus making it more visible. In addition to the journals, the monographic production is included in DAIS, and it contains more than 60 volumes. This part of the first phase was finished and the Institute presented a retrospective of its publishing activity in DAIS (250 volumes of magazines and monographs).

The second segment in the first phase of the establishment of DAIS was the collection, preparation and distribution of articles of prominent ISJ SANU scholars. In the first year, more than 1,200 scholarly articles entered the repository.[12] A small number of these articles already existed in digital form and the whole digitization work was supposed to go through the process of physical preparation and work on scanning, editing scanned documents, and creating metadata. According to Foster and Gibson, the institutional repository "without content, is the same as a series of empty

[10] In order to make our documents readable we reduced the size of the files (because some also had more than 1 GB). The size of the document burdens the repository and its resources as well as the end user who wants to get an optimal size document in a short time. For example, a year's magazine has an average of 200 and 300 Mb, but after Google's optical character recognition, the document has been reduced to approximately 30 Mb.

[11] So far, the evaluation of the success of optical text recognition has not been done. When services of a domestic initiative that deals with the optical text recognition, especially the Cyrillic are available and affordable, we will gladly re-enter the OCR and repossess the documents with more precisely recognized text into the repository.

[12] They include Academic Irena Grickat, Academic Mitar Pešikan, Dr. Ivan Popović, Dr. Berislav Nikolić, Scientific adviser dr. Egon Fekete, Prof. dr Dimitrije E. Stefanović, Academic Aleksandar Loma, Scientific advisor dr. Jasna Vlajić-Popović, Scientific advisor dr. Stana Ristić, Scientific advisor dr. Sreto Tanasić and others.

shelves" (Gibbons and Foster, 2005, par 4).[13] Therefore, the establishment of user habits and setting a good example is essential to empower the initial dynamics of a repository.

Creating and organizing digital content in DAIS is in line with the requirements of the European Commission concerning open access to publications, and it enables the further dissemination of scientific information through European and international portals such as OpenAire, BASE, CORE and Google Scholar. One of the fields for entering metadata is the name of the project within which the work was created. This field, through the link, connects all the works of a particular project to DAIS, but also in the OpenAire database, which in one place collects data of the results of all projects of the European Commission, as well as data on the results of national projects from about ten countries, one among them being Serbia (OpenAire, 2019). Also, using the ORCID (Open Researcher and Contributor ID) identifier, all of the author's publications are combined and linked in one place, and published under different variant names.

When determining the profiles of the digital collection, ISL SANU decided to provide full text in open access for all of its editions, as well as for the published work of the employees, with restrictions that determine the copyright of other publishers. As for the ISL editions, it was decided to apply CC licenses. This simple and standardized way of copyright regulation has multiple levels. The institute opted for the CC BY-NC-ND module (Attribution – NonCommercial – NoDerivatives), which implies that it is mandatory to indicate the name of the author and that the author gives permission to download and further distribute the work. This most restrictive of free licenses does not allow further processing of the work or its commercial use.

In its architecture, DSpace software allows creating collections in the digital repository. These collections categorize and sort content thematically, formally or otherwise. At present, ISL SASA in DAIS has seven collections. In addition to the four collections that classify the periodicals of the Institute, the two are created for monographic series, while the remaining one is general. The general collection collects the Institute's Proceedings and other materials, as well as the work of researchers published outside the Institute. The problem of classification always imposes various solutions. The material

[13] As an example of good practice, we can mention the repository of the Massachusetts Institute of Technology (MIT), which has owned over 106,000 digital records in its repository by 2019 (DSpace, 2019)

could have been classified differently, but the optimal approach also takes into account the easy navigation by the end users.

4 Conclusion

During the relatively short period of time in the field of digital humanities, ISL SASA has undergone major changes. The institutional repository has come to life and serves a number of purposes. The main purpose is to promote the concept of Open Science and to represent the scientific production of the Institute, which fulfills the criteria provided by the Ministry's Platform. What is publicly funded should be publicly available. The institutional repository is also developing the presentation of retrospective scholarly production. Almost everything published by ISJ SANU has been put on the repository. This way the old scholarly papers that are still relevant become easily accessible. The library has thus become a participant of scholarly research. According to Baudoin and Branschofsky, the creation of a digital repository changes the way we think about the life cycle of scientific research (Baudoin and Branschofsky, 2004, 43).

Thanks to the successful cooperation of the library and the management of the Institute, as well as research associates, the digital repository is established, filled with scholarly papers and ready to follow future scientific production. Through the successful cooperation of all factors in this complex process, the credibility of the established repository is being built and promoted. It is important to note that the ISJ researchers have shown willingness to engage in the work, which had a significant influence on the results. Now the researchers' production is almost completely in free access. The percentage of open access content in the ISL SANU collections is 100%, which is a great success, bearing in mind that the open approach to humanities in the world is much less represented and develops more slowly than in other disciplines. According to reports made for the needs of the European Commission, the greatest restriction in open access is within social sciences and humanities (Archambault and Roberge, 2014, 20).

References

"DSpace". Duraspace. Accessed: 12.03.2019, https://duraspace.org/dspace/about/

"OpenAire". European Union/European Commission. Accessed: 08.03.2019, https://www.openaire.eu

"Платформа за отворену науку". Министарство просвете, науке и технолошког развоја. Accessed: 21.02.2018, http://www.mpn.gov.rs/wp-content/uploads/2018/07/Platforma-za-otvorenu-nauku.pdf

Archambault, É., D. Amyot, P. Deschamps, A. Nicol, F. Provencher, L. Rebout, and G. Roberge. "Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels – 1996–2013". *Science-Metrix* (2014). Accessed: 20.03.2019, http://science-metrix.com/en/publications/reports/proportion-of-open-access-papers-published-in-peer-reviewed-journals-at-the

Baudoin, Patsy and Margret Branschofsky. "Implementing an Institutional Repository: The DSpace Experience at MIT". *Science & Technology Libraries* Vol. 24, no. 1/2 (2004): 31–45.

Branin, Joseph. *Encyclopedia of Library and Information Science*. Institutional Repositories. New York, N.Y.: Marcel Dekker, 2004-2005.

Ferreras-Fernández, Tránsito, Francisco J. García-Peñalvo and José A. Merlo-Vega. "Open Access Repositories as Channel of Publication Scientific Grey Literature". In *TEEM '15 Proceedings of the* 3^{rd} *International Conference on Technological Ecosystems for Enhancing Multiculturality*, 419–426. New York, NY, USA: ACM, 2015.

Gibbons, Susan and Nancy Fried Foster. "Understanding Faculty to Improve Content Recruitment for Institutional Repositories". *D-Lib Magazine* Vol. 11, no. 1 (2005). Accessed: 25.03.2019, http://www.dlib.org/dlib/january05/foster/01foster.html

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West and Stefanie Haustein. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles". *PeerJ* Vol. 6 (2018). Accessed: 01.04.2019, https://peerj.com/articles/4375/;6:e4375, DOI:10.7717/peerj.4375

Rajović, Vasilije, Biljana Kosanović and Milica Ševkušić. "DSpace – institutional repositories – dissemination of research results: A local case study". In *Primena slobodnog softvera i otvorenog hardvera PSSOH 2018*. Belgrade, Serbia: University of Belgrade – School of Electrical Engineering and Academic Mind, 2018.

Van de Sompel, Herbert and Michael L. Nelson. "Reminiscing About 15 Years of Interoperability Efforts". *D-Lib Magazine* Vol. 21, no. 11/12 (2015). Accessed: 15.03.2019, http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html

Winter, Marsha and Portia Bowen-Chang. "Dealing with DSpace: The Experience at the University of the West Indies, St. Augustine". *New Library World* Vol. 111, no. 7/8 (2010): 320–332.

Ćirković, S. "Grey Literature – The Chameleon of Information Resources". *Infotheca - Journal For Digital Humanities*, Vol. 18, no. 1 (2018): 75–83. doi:10.18485/infotheca.2018.18.1.5

Шевкушић, Милица, "Партнерски програм Гугл књиге као платформа отвореног приступа у научним библиотекама". In *Отворен приступ знању у библиотекама, организатори конференције Библиотекарско друштво Србије [и] Народна библиотека Србије*, 187–207. Београд: Библиотекарско друштво Србије, 2013.

Digital Library "The Great War" – development and results

UDC 027.54(497.111):004.738.5
DOI 10.18485/infotheca.2019.19.1.5

ABSTRACT: While working on the project *Europeana collections 1914–1918*, the expert team of the National Library of Serbia made a decision to create a separate, thematic digital library, which would contain materials from the period of the World War I related to Serbia and the Serbian people. The paper gives an overview of the process and all the stages involved in the development of the portal and digital library *The Great War*. With the help of Google Analytics service, conclusions and basic statistics on the use and users of this digital library were presented.
KEYWORDS: digital library, The Great War, the First World War, National Library of Serbia, development, users, statistics.

PAPER SUBMITTED: 15 April 2019
PAPER ACCEPTED: 29 May 2019

Biljana Kalezić
biljana.kalezic@nb.rs
National Library of Serbia

1 Introduction

On the occasion of marking the centennial since the beginning of the First World War, the International Foundation *Europeana*[1] launched several projects in 2012: *Europeana collections 1914–1918*,[2] *Europeana* 1914–1918[3] and EFG1914.[4] All of these projects had a common goal – to digitize publications and objects from the 1914–1918 period and make them available to public, as well as materials from a later period that are thematically related to the First World War (Калезић, 2014).

[1] International Foundation Europeana (accessed on 27/03/2019)
[2] Europeana collections (accessed on 27/03/2019)
[3] Europeana (accessed on 27/03/2019)
[4] European Film Gateway 1914 (accessed on 27/03/2019)

The basic idea of the *Europeana collections 1914-1918* project was to digitize and make available online materials from the collections of nine national libraries of the countries that participated in this historical conflict on different sides. The library material from this period, until this moment available for use only in reading rooms of these libraries, is often in very poor physical condition concerning its age and quality of production, so their digitization greatly contributed to preservation and protection of this sensitive material. In this way, over 400,000 publications have been digitized and preserved for the future, and have been given free access at the website of the project. The National Library of Serbia is one of the twelve equal partners in this project. The creation of the portal *The Great War*,[5] with the digital library being its most significant part, was one of the results of participation of the NLS in the project. According to the project guidelines, all of the materials originating from the First World War period related to the Kingdom of Serbia and the Serbian people were included in this digital library. Development of the portal began in November 2012, and the first version was installed on January 5, 2013. The public promotion was held on the Day of the National Library of Serbia, February 28, 2013.

2 The Great War Digital Library – selection of material

The types of material included in the digital library *The Great War* are defined by the *Europeana collections 1914-1918* project guidelines: books, periodicals, war diaries, printed music, children's literature, photographs, posters, propaganda materials and everything else that can be found in the collections of national libraries and their digital counterparts. However, the selection of material to be digitized and included in the digital library *The Great War* was largely conditioned by the specific position of the Kingdom of Serbia and the Serbian people in the First World War.

At the beginning, the 20^{th} century publishing in Serbia was largely undeveloped, with the majority of the material, nearly 80% of it, being published in Belgrade. During the First World War, at the time when Belgrade became a front line and remained so for more than a year, Serbian publishing had suffered an additional blow because of the general scarcity. However, the occupation of Serbia in 1915, the retreat of the army and state apparatus through Albania to Greece, and later the dispersal of refugees across

[5] The Great War (accessed on 10/04/2019)

Western Europe and the world led to the opening of new centers of publishing activity of the Serbian people. Serbian printing offices in Corfu and Thessaloniki were established, while the printing office of Serbian military invalids in Bizerte in Tunisia was especially active. Scientific editions and school leaflets in Serbian or languages of the host countries appeared at university centers of Western European countries (France, Great Britain, Italy, Switzerland), while pro-Yugoslav emigration in South and North America published a number of publications devoted to the war and ideas of the establishment of the Yugoslav territory after the war. At the same time, , the Austro-Hungarian and Bulgarian authorities were publishing their newspapers, proclamations to the public, etc. in occupied Serbia. A large number of propaganda brochures were published in the Austro-Hungarian Monarchy with a purpose of strengthening loyalty of the Slavs, and especially the Serbian population to the Monarchy (Калезић, 2014).

Bearing all historical circumstances in mind, for the selection of materials and creating this digital library, it was necessary to form a multidisciplinary team composed of librarians, historians, IT professionals, history teachers and professors. The set goals were to make the material as comprehensive as possible, but also to present it in such a way that the search and use of the material is easy, whether for a beginner or more advanced digital library users. The comprehensiveness of the material could not be achieved if there was no cooperation with other institutions from the country and worldwide, which by digitizing the material from their own collections significantly enriched this digital collection. Some of the institutions with which the National Library of Serbia had cooperated are: the Archives of Serbia, the National and University Library of the Republic of Srpska, the Military Museum, the National Assembly of the Republic of Serbia and many others (Калезић, 2014).

In the possession of the Special Collections Department of the National Library of Serbia, apart from the mentioned material, a large number of diaries, letters, photographs and other materials are kept, which provide insight into the daily life of the occupied population, as well as soldiers at the front, wounded in hospitals and prisoners. By incorporating these materials into the digital library *The Great War* the goal set by the expert team of the project was fulfilled – to give users an opportunity to comprehend and understand many social processes in given period, that significantly changed the image of Serbia, Europe, and even the world. Accordingly, and taking into account the needs of future users, it was decided that the material

important for the study of this period which was published after the First World War until 1941, should also be included.

3 Creating the collection

After the selection of materials was done, the work on its digitization has started. Employees from the Department of Digitization of the National Library of Serbia started this task. As the scope of work and the pressure to finish digitization in a certain period of time grew, it became clear that help from the outside was necessary. Shortly thereafter, a contract was signed with the *Ebart Media Archive*, which as the subcontractor continued the digitization process. It was agreed that the resolution of the scanned material should be 300 dpi. Apart from material from the National Library of Serbia, materials from partner institutions that were not eligible for digitization were also scanned by *Ebart*. Certain institutions that were able to do so according to the guidelines they received from the National Library of Serbia, delivered finished digital copies. Work on digitization lasted for about two years. Today, the digital collection of *The Great War* contains 9,698 documents of all categories of material, on 96,473 pages.

For the entry and display of digitized material, the open access platform of Omeka v. 1.5.3[6] was chosen. The applied metadata scheme is Dublin Core Extended.[7] Each entered item has an appropriate description, which is linked to the bibliographic description in the local database of the National Library of Serbia via the COBISS ID number. Metadata is partially downloaded to the system, and partially manually entered. The advantage of this method is that the automatically downloaded descriptions and data already passed bibliographic control, which reduced the possibility of incorrect and unverified information. The material that was added to the digital library on the basis of cooperation with other institutions was also described in accordance with the current standards (Калезић и Михаиловић, 2014). The entry of the material that consists of a large number of connected digital objects, such as books or periodicals, was facilitated with the installation of Dropbox app. Namely the technical capabilities of the installed version of Omeka did not allow the simultaneous entry of multiple digitized pages, but they had to be entered separately. By installing the Dropbox it became possible to enter a

[6] Omeka (accessed on 23/05/2019)
[7] Dublin Core Extended (accessed on 23/05/2019)

complete folder – with all the digitized pages of a single document to Omeka at the same time.

For periodical publications, the metadata scheme was specifically developed. The basic description for the individual title was taken from the COBISS system, but for a description of individual issues, additional information had to be entered. So in this case, the date of the individual issue was entered next to the title, and by filling in additional fields in the metadata scheme, more detailed identification and connection with the basic description for the entire title was enabled. Thus, the following fields were filled in: *Description* – with a year and an issue number, *Date* – with a date, and the field *Is part of* – with a link that leads to a description of the entire title, i.e. collection.

Due to the fact that the calendar application could not be installed, and with the aim of further simplifying the later search, the Omeka *Tags* were used for a month and a year of a publication, using the following structure: the Roman numerals were entered for a month, and the Arabics for a year. With these tags, it became possible to form a segment *Available material* next to a description of an individual title, which will be discussed later.

Metadata for cartographic materials is additionally enriched (Glišović and Gardašević, 2015). In order to overcome the language barrier and provide users with additional information about digitized maps, it was decided to enter a URI (Uniform Resource Identifier) in the field named Spatial Coverage. In this case, as a URI, a link to the geographic database GeoNames[8] is used. By clicking on a link, you get information about particular geographic location, e.g. population, region, geographic coordinates, different versions of names, etc. Due to various limitations, it is not possible to use all versions of geographic location names from the GeoNames database in a search of cartographic materials, but for most toponyms, the *Alternative Title* field has been filled in with English version, thus facilitating a search for users who do not know the Serbian language (Glišović and Gardašević, 2015).

At the time of writing the project, which was the end of 2012, the technical capabilities of the Omeka program did not allow the full-text search of the documents. Having these deficiencies in mind, the expert team had made efforts in these ways to overcome this shortcoming of Omeka.

[8] GeoNames (accessed on 10/04/2019)

4 Search and display of results

All of the items in the digital library *The Great War* have been sorted into collections for easier navigation and more transparent data display. Formal and content principles were used to form collections. The formal one followed the basic division of materials according to their type, but the content one proved to be necessary. In this way, the material is divided into the following collections: books, periodicals, manuscripts, posters, paintings, cartographic materials, and various items. However, members of the expert team, based on their experience in working with other digital libraries, and also through research of historical material, considered it useful to form sub-collections, because the collections would be more difficult to search if the formal principle of material distribution remained the only one applied. So, for example, in the collection Books, there is a division into several sub-collections, some of which are: *Literature, Legal Regulations, The Great War of Serbia, The Progress Library.* This has been done with the intention of grouping related publications in one place and providing users with easier navigation and comparative analysis of individual titles.

The title screen contains the keyword search. Also, users can independently navigate through specific collections. Next to the individual titles of periodicals, on the right side of the screen there is a segment *Available material*, where users can browse through the years and months, and the material from the specific period of time can be directly accessed.

However, thanks to the advanced search capabilities, users can significantly narrow search criteria, practically by all elements of a description of a digitized object, such as year of publication, author, type of material, etc. Further narrowing of a search can be done using collection filters or the aforementioned *Tags.* To search using the *Tag*, one should take into account that query format must match the structure of these tags, so when searching, the month mark must be entered in Roman numerals, the year in Arabic and they must be separated by comma, for example: II, 1915.

As a result of a search, a page with results that match the set parameters is shown, with a small image, a title and a year of publication of an item. By clicking on any of these two, a page with a detailed description of the requested publication and a reduced display of the first page of digitized material opens. By clicking on a picture, a digital object opens further and, depending on a type of material, it is possible to continue to scroll through with the arrows, but also to skip to the desired page. The material can be zoomed in, and the right click allows a page to be saved as an image on the

user's computer. General conditions for usage of files downloaded from *The Great War digital library*[9] are defined in the *Terms of Use* section.

5 Additional content

Since the very beginning, there was a plan for *The Great War* to be used as an auxiliary tool in schools. The idea to form a section *Learning*, along with the digital library, was born thanks to some members of the expert team who had a long-standing experience in education. Through this segment, teachers and students are able to find and use materials related to the First World War as a teaching tool. Teachers can find pedagogical advice, educational standards for the primary education in the field of history, recommended materials on selected topics, and preparation for classes with links to material on the portals *The Great War* and *Europeana*. Although it is closely related to the First World War, this digital library can, in addition to its apparent application in the teaching of history, have its application in the teaching of Serbian language and literature, religious education, fine arts, and other topics (Ковачевић и др., 2016).

In order to further promote this segment of *The Great War* portal, the seminar "Digital Libraries dedicated to the First World War as a teaching tool" was devised. The seminar was accredited by the Institute for the Advancement of Education and was held on five occasions during 2014 and 2015. Attendees of the seminar were mostly teachers and professors of history at primary and secondary schools. During the seminar, they got acquainted with the collections and search capabilities of the most important digital libraries dedicated to the First World War, with a special focus on *The Great War*, and as well as with ways to bring these contents closer to students and encourage their curiosity and further research.

Virtual exhibitions are yet another segment through which the materials from the digital library are presented to users. It provides them an opportunity to get acquainted with events, themes and people that have left their marks during this period. The current exhibition *Crnjanski in the War*, for the time being, has been the only one.

There is also a section *Timeline* on the home page, in which the material is automatically and randomly displayed on the appropriate date.

The interactive map is another valuable segment of the portal. On this world map users can, through geo-visualization, find a place of publication

[9] General conditions of use (accessed on 27.3.2019)

of specific materials published during and immediately after the war. An interactive map was created on the Google platform. By selecting a region or a city, direct links to the publications are shown. Using the filters for a year of publication, the type of material and the language of a publication, search through the map can be further narrowed (Калезић и Михаиловић, 2014).

6 Results

After just over six years of existence and operation of the digital library *The Great War*, it is possible to analyze the statistics on users and usage, as well as achieved results, thanks to the Google Analytics[10] tool. Google Analytics allows you to track the number of users, their geographic location, length of each visit, number of pages that users viewed in one visit, and various other parameters.

So, from the beginning of *The Great War* promotion until December 31, 2018, the total number of visits was 102,550. The total number of individual users was 58,588, out of which 17,7% were returning users. During this period, a total of 2,081,557 pages were viewed, i.e. just over 20 pages per visit. The average duration of one visit was about six minutes.

Table 1 gives an overview of these data by year, since 28^{th} February 2013, to the end of 2018.

Year	Number of Users	Number of Visits	Number of Pages viewed (per visit)	Average Duration of Visit (minutes)
2013	7.169	14.571	295.229	6,01
2014	11.797	24.465	657.023	8,09
2015	10.792	18.222	348.851	5,37
2016	10.338	15.605	274.453	5,07
2017	7.635	12.921	222.684	5,04
2018	10.854	16.766	283.317	4,46

Table 1. Data on the usage of the digital library *The Great War*

[10] Google Analytics (accessed on 10.4.2019)

In the first three months of 2019, 2,238 users (17.9% of them returning) made 3,332 individual visits and viewed 58,763 pages, with an average length per visit of 5.16 minutes.

Figure 1. Geographic location of the users (28/03/2013 - 31/03/2019)

According to the geographic location of the users, as expected, most of them were from Serbia and countries in the region. But among the top 10 countries by the number of users, other countries that were the main participants in the First World War on both sides were also present, according to Google Analytics data.

Other demographic data that can be analyzed is the age and gender of users. Out of the total number of users, 54.15% were male and 45.85% female. As for the age, most of them, 33.5%, belong to a group of people between 25 and 34 years old, followed by users aged 18 to 24 - 27.5%. The share of users aged 35-44 is 15.5%, those from 45-54 was 12.5%, from 55 to 64 was and 5.5%, and the same for those over 65 years of age.

Google Analytics is also tracking device data used to access The Great War. Most visits were from desktop computers, i.e. about 75% of all visits, about 22% were from a mobile phone and about 3% of the visitors used a tablet.

According to Google Analytics, the data on the use of materials in the period 28/03/2013 – 31/03/2019, showed that the most visited document was the monograph *War Album 1914–1918* by Andra Popović, published in 1926 with 10,087 individual visits. The next were the first and the fourth

Country	Number of Users	Number of Visits	Number of Pages viewed (per visit)	Average Duration of Visit (minutes)
Serbia	41.524	79.551	21,36	6,12
BiH	3.017	4.075	21,08	6,32
USA	1.552	1.954	8,16	2,20
Montenegro	1.103	1.752	21,43	6,33
Croatia	1.059	2.135	24,78	10,03
Germany	903	1.179	13,62	3,39
Bulgaria	888	1.158	19,64	4,40
France	724	866	9,68	2,58
Great Britain	707	819	8,75	2,35
Macedonia	704	1068	18,94	5,34

Table 2. Overview of the number of users by geographic location and data on the use of digital library The Great War (28/02/2013 – 31/03/2019) – the first 10 places

volumes of the *The Great War of Serbia for the Liberation and Unification of Serbs, Croats and Slovenes* collection with 8,651 and 4,779 visits. This collection, published from 1924 to 1937, in which the documents of the Supreme Command of the Kingdom of Serbia were published, is an invaluable source for anyone who is researching the period of the World War I, and therefore it is not surprising that the number of views of individual books is high.

Among the frequently visited pages was the sticker album – *World War in Photos: 1914 – 1918: Milk chocolate Šonda* from the 1930s, which had 4,165 visits. Of the periodicals, the fifth volume of the magazine *Woman* from 1918 was the most visited – 3,661 visits. The postcard from France – *Serbie: la revanche ou la mort* has 3,071 visits, and the postcard with the description "Momčilo Gavrić, 10 years old, whose parents were killed by Austrian soldiers, is reporting with his friend to the officer" – 2,987 visits.

The monograph *Tragic Days of Belgrade*, by Jovan Miodragović from 1915, is the most visited item published during the war period with the total of 2,953 visits.

7 Conclusion

The Analyzed user data and the use of *The Great War* portal show that this digital library, with an average number of monthly visits of around 900, is a very valuable and important source for studying the period of the First World War. The users' age structure shows that more and more people give preference to easily accessible and searchable digital sources of information compared to traditional ones. All this points to a need for the expert team to continue to work responsibly on the development and promotion of *The Great War* portal. Although the basic project is completed, the work on improvement and enrichment continues, by continuously adding scanned materials, and by holding lectures where in direct contact with the users, both returning and new ones, the team gains an insight and new ideas for the future of the portal.

References

Glišović, Jelena and Stanislava Gardašević. "Cartographic Collection of National Library of Serbia throughout History until the Digital Present". *e-Perimetron* Vol. 10, no. 2 (2015): 73–86. Преузето 01.04.2019, http://www.e-perimetron.org/Vol_10_2/Glisovic_Gardasevic.pdf

Калезић, Немања. "Дигитална библиотека 'Велики рат'". Српске студије". Vol. 5 (2014): 389–391.

Калезић, Немања и Наташа Михаиловић. "Народна библиотека Србије, Интерактивни портал Велики рат". Нови Сад: Заједница матичних библиотека Србије, Градска библиотека, 2014, 147–150.

Ковачевић, Огњен, Наташа Михаиловић и Катарина Стаменић-Станојевић. *Приручник за примену дигитализованих историјских извора у настави*. Београд: Народна библиотека Србије, 2016.

Audience in Focus – Democratisation of Digitalisation in Libraries

Jelena Andonovski
andonovski@unilib.rs

Nikola Krsmanović
krsmanovic@ubsm.rs

University Library "Svetozar Marković" Belgrade, Serbia

PAPER SUBMITTED: 5 June 2019
PAPER ACCEPTED: 8 July 2019

In the most general sense, digitalization has allowed for permanent preservation of only a small part of the material of cultural heritage, both in the world and in Serbia. An even smaller percentage of this kind of stored material has been processed and therefore effectively available to users. In recent years, the institutions have controlled the processes of digitization defining the flow of information created based on the use of digital materials of cultural heritage. Today's technological capabilities enable individuals to control digitalization processes, opening up a number of issues related to digitization of books and cultural heritage. University Library "Svetozar Marković" in Belgrade is an associate partner of the H2020 READ (Recognition and Enrichment of Archival Documents)[1] project team, which is implemented from 2016 to June 2019. The core objective of READ is to provide a service platform for the automated recognition, transcription and searching of historical documents, Transkribus[2]. The main objective of Transkribus is to support users who are engaged in the transcription of printed or handwritten documents, namely humanities scholars, archives, members of the public and computer scientists. The basis and main precondition for the use of Transkribus services are digitized documents. In order to simplify and speed up the digitization process, researchers at the Computer Vision Lab Laboratory at the Vienna University of Technology[3] are developing a mobile DocScan

[1] Recognition and Enrichment of Archival Documents, on-line
[2] Transkribus, on-line
[3] TU Wien, Faculty of Informatics, Institute of Visual Computing & Human-Centered Technology, Computer Vision Lab, on-line

application and a portable ScanTent device. The DocScan is an Android app designed for the ScanTent. It detects pages in the live preview and makes high-quality scans. An automatic series mode takes an image once a page is turned. It therefore enables you to scan books or documents quickly without interacting with your mobile. DocScan is an open source and is available on the GitHub platform, as well as on the Google Play Store. The ScanTent is a portable piece of equipment that holds a smartphone in place above a document. In combination with the DocScan app it enables users to hold a document with both hands and to scan it with your smartphone without pressing any button.

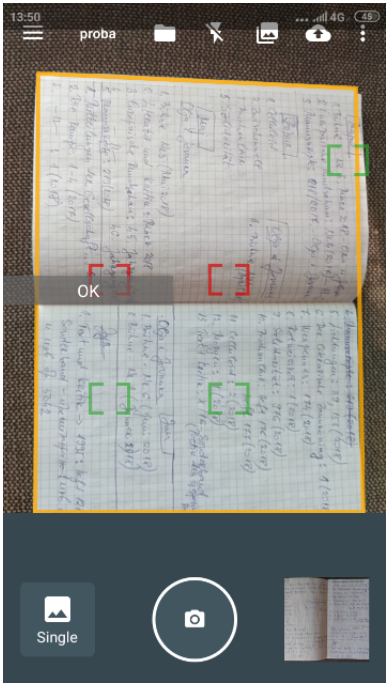


Figure 1. DocScan app

.

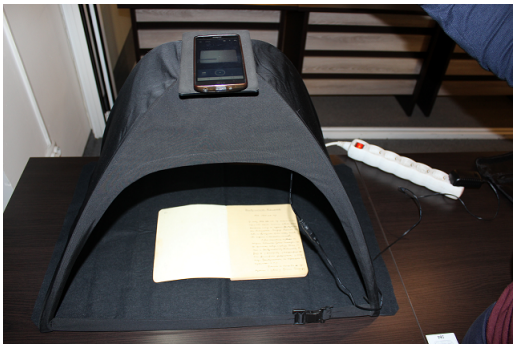


Figure 2. ScanTent device

.

Parts of the H2020 READ program were realized in the University Library "Svetozar Markovic" in Belgrade, within the framework of projects financed by the Ministry of Culture and Information of the Republic of Serbia: "The New Horizon of Digitization" for 2016, "The Ready Old Serbian Cyrillic: Hand-written Handwritten History" 2017, "The Readiness of the Old Serbian Cyrillic: History and Tradition at Your Fingertips" for 2018. Also, in cooperation with the City Library of Novi Sad it was organized project "Democratization of Digitization in Libraries" supported through the "Audience in Focus" open call of the Foundation "Novi Sad 2021" with the funds allocated from the budget of the Autonomous Province of Vojvodina for 2017.[4] The aim of the project was to train librarians, as well as to involve as many users as possible in the process of digitization of library resources in order to make library and archive resources more available. During October, November and December in 2018, as well as in March, April and May in 2019, it was held more than 30 workshops in order to present the new technological possibilities based on digitization. The workshops were held in the University Library in Belgrade and the City Library of Novi Sad by the team of the University Library for the development and improvement of new technologies in the field of digitization. In December 2018 (20^{th} and 21^{st}), were held two national workshops under the patronage of the Ministry of Culture and Information of the Republic of Serbia in the University Library in Belgrade and the Historical Archive of Novi Sad. Workshops were led by Dr Ginter

[4] Democratisation of Digitalisation in Libraries, on-line and on-line

Milberger, the leader of the pan-European team for the development of new technologies in digitization.

Figure 3. The National workshop in Belgrade

.

Figure 4. The National workshop in Novi Sad

.

Further work and training of the mentioned technologies continues in the University Librar "Svetozar Marković" in Belgrade within the project "Luxury of fully read Serbian Cyrillic: living manuscript heritage from Branković to Obrenović" financed by the Ministry of Culture and Information of the

Republic of Serbia for 2019, and the accredited program for professional development "Democratization of Digitization in Libraries" by the authors Nataša Dakić, Aleksandra Trtovac and Jelena Andonovski. After completion of the H2020 READ project, a pan-European Collaborative will be established to ensure the long-term sustainability of the project's results.

Implementation of authority control in COBISS.SR

Svetlana Pucarević
svetlana.pucarevic
@unilib.bg.ac.rs

Snježana Furundžić
snjezana.furundzic
@unilib.bg.ac.rs

University Library
"Svetozar Marković"
Belgrade, Serbia

PAPER SUBMITTED: 6 May 2019
PAPER ACCEPTED: 17 May 2019

After being implemented in COBISS.SI (Slovenia) and COBISS.BG (Bulgaria), authority control was implemented into COBISS.SR (Serbia). At the moment, the system COBISS.SR includes only the personal names authority file of authors, but there is a plan in the second phase to include corporate names, as well as the subject headings authority file in some further work. The personal names authority file in the domestic system is called CONOR.SR. Bibliographic and authority records are linked through the CONOR.SR database. CONOR.SR is a unique database for all Serbian library members of the system COBISS.SR which contains unique authority record for each author's name that appears in the bibliographic record. CONOR.SR database project was realized by the Matica Srpska Library, the National Library of Serbia and the University Library "Svetozar Marković", with the participation of all libraries in Serbia who maintain and develop their cataloguers under the COBISS platform. The initial CONOR.SR database was installed in 2013 and edited in the period from 2013 to 2018. A complete implementation of authority control in the system COBISS.SR included the following steps:

One-day course for cataloguers

During 2018 librarians from the three libraries mentioned above were trained to educate cataloguers from libraries how to use the system COBISS.SR. The educators from the University Library "Svetozar Markovic" held 12 courses

for about 250 cataloguers from the University libraries in Belgrade, Niš and Kragujevac. After that all cataloguers got privileges for creating basic records in the authority file.

Course for editors of authority records

The National Library of Serbia organized a training for updating data in the authority file records intended for the so-called "Middle" editors, who were given more privileges. The most privileges were given to the so-called "Top" editors.

Update of existing CONOR.SR database

The year 2018 and the beginning of 2019 were key periods in terms of editing records and deleting duplicate records in the existing CONOR database. Particular attention was paid to identifying and deleting duplicate records because after the implementation of CONOR database, this cannot be possible. As bibliographic records are linked with appropriate authority records, deleting an authority record would mean that a specific bibliographic record would be left without an author heading (authority access point) and would not be correct in OPAC representation.

In this period, it was important to check and correct as many records as possible of personal names that are significant to our culture, as well as records of authors whose names are associated with the largest number of bibliographic records. The University Library "Svetozar Markovic" has the responsibility to edit the authority records of local researchers, whose names are linked with the largest number of bibliographic records in COBISS.

Linking authority and bibliographic records

A test CONOR database was established in early April 2019 and enabled cataloguers to become familiar with the system and to exercise the work process before installing the real database. During the weekend of April 13th and 14th, 2019, IZUM (Institute of Information Sciences) installed the software for authority control and connected CONOR to COBIB and local databases. The CONOR database has been active in our COBISS.SR system from April 15th, 2019. When creating a record, there is no longer a manual entry of the author heading in block 7XX. There are two possibilities to fill block 7XX:

– to link with the corresponding authority record if it already exists in the CONOR database
– or to create a new one if there is no appropriate authority record in the CONOR database.

As Serbian libraries have specific situation regarding to a parallel use of the Cyrillic and Latin script, the creation of the uniform heading for the personal name means that there are two parallel fields 200 in the one CONOR.SR file. The first authority access point is always in the Cyrillic, and the second one in the Latin. Some examples illustrate this:

Example 1. – A domestic author

200 1 <7>cb - cyrillic - not specified <a>Црњански<b>Милош<f>1893-1977
200 1 <7>ba – latin<a>Crnjanski< b>Miloš <f>1893-1977

Example 2 – A foreign author whose name is originally written in Latin

200 1 <7>cb - cyrillic – not specified <a>Ирби<b>Аделина Паулина< f> 1831-1911 200 1 <7>ba – latin<a> Irby <b> Adelina Paulina <f> 1831-1911

Example 3 - A foreign author whose name is originally written in Russian Cyrillic

200 1 <7>ca - cyrillic - not specified <a>Достоевский<b>Фёдор Михайлович< f>1821-1881 200 1 <7>ba - latin<a>Dostoevskij<b>Fedor Mihajlovič<f>1821-1881

For Russian authors, it is also necessary to enter the name reference in the field 400 which is an alternative access point. For example:

400 1 <7>cb – cyrillic - not specified <a>Достојевски<b>Фјодор Михајилович<f>1821-1881

At the moment of writing the paper, the number of authority records in the CONOR.SR database was about 52000.[1]

[1] The data was retrieved from the CONOR database on April 25, 2019.

Full synchronization of the system and linking of bibliographic records created before April 15th, 2019 to CONOR will be done by the Institute of Information Sciences (IZUM) after May 1st, 2019.

Working with authority records is a precondition for linking data from local and union catalogues to Linked Open Data. We believe that we will be successful in the new work process and that authority control will improve both the quality of bibliographic records and their uniformity.

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).