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, University of Belgrade

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

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

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

Belgrade, Vol. 21, No. 1, September 2021

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

+381 11 3370–211
infothecajdh@gmail.com

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Contents

FrameNet Lexical Database: Presenting a Few Frames Within the Risk Domain

UDC 81'322.2
DOI 10.18485/infotheca.2021.21.1.1

ABSTRACT: This paper gives a short overview of the frame semantics theory that forms the theoretical basis of the Berkeley *FrameNet* project. We present the basic concepts of this database, as well as the possibility of implementing it in Serbian. We also take a close look at the lexical analysis used in the FrameNet development project and point out the differences between the frame-based lexical analysis and its word-based counterpart. This is followed by an illustration of a couple of related frames evoked by words from the risk domain. FrameNet data is also readily available through the Python API included in the NLTK (*Natural Language Toolkit*) suite, which provides a good natural language processing resource. The last chapter shows a corpus search of the noun *risk* in a mining-themed corpus. We also present its most common collocates, word sketch, individual pattern concordances, thesaurus entry of its synonyms and related words, collocation frequency graphs. A word cloud for the word *risk* is also included.

KEYWORDS: Serbian language, frame semantics, FrameNet, risk scenario, mining corpus, natural language processing.

PAPER SUBMITTED: 15 July 2021
PAPER ACCEPTED: 6 September 2021

Aleksandra Marković
aleksandra.markovic@isj.sanu.ac.rs
Institute for Serbian Language, SASA
Belgrade, Serbia

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

Natalija Tomić
ntomic@hotmail.com
University of Belgrade
Faculty for Mining and Geology

Olivera Kitanović
olivera.kitanovic@rgf.bg.ac.rs
University of Belgrade
Faculty for Mining and Geology
Belgrade, Serbia

1 Introduction

Charles Fillmore's Frame Semantics Theory is a cognitive theory of meaning that links word meanings to the syntactic context in which they occur (Atkins, Fillmore, and Johnson 2003, 254). Word sense analysis is tra-

ditionally left to lexicographers and those interested in semantics. However, if the aim is to show the manner in which a word is actually used, an analysis of corpus data proves to be a fairly complicated task, in view of the number of concordances proposed by contemporary corpora for certain key words. Frame semantics theory, as cited by the following authors (Atkins 1994; Gildea and Jurafsky 2002; Atkins, Fillmore, and Johnson 2003; Pradhan et al. 2005; Boas and Dux 2017; Jurafsky and Martin 2020), gives a reliable, scientifically valid way of approaching word usage analysis and description. The basis of this approach is the idea that every experience that we memorize occurs in some meaningful context and our ability to memorize those experiences stems from the existence of mental schemas that we possess giving meaning to objects, relationships and events. Fillmore argues that words are learned within such meaningful contexts, and that context is also essential to the process of comprehension, when we evoke specific experiences through which we learned the meaning of a word. A frame identifies the type of experience and provides its structure and coherence, lending meaning to entities, events and relations that make it up (Fillmore 1976, 26).[1]

1.1 The design of FrameNet

FrameNet[2] is a lexical database of English based on annotated examples of how a *lexical unit* (hereinafter abbreviated as LU) is used in an actual7 texts. The basic premise comes down to the fact that most LUs are best defined through semantic frames, a conceptual structure that provides a description of the type of situation, relation or entity and the participants involved in it (Ruppenhofer et al. 2016, 7). For example, taking a risk typically involves the following: a person taking the risk that is central to the RISK scenario or the Protagonist. The Protagonist takes a risk willingly or otherwise or runs the risk; possible Bad outcome or Harmful event; the Decision which may lead to a bad outcome; a Purpose; an Action; certain Circumstances in which the protagonist stands; an Asset (a person or an object), perceived by the Protagonist as desirable, all of which is compromised in the RISK scenario (Fillmore and S. Atkins 1994, 367).

1. The term *frame* in Fillmore's usage denotes a general signifier that can be referred to as schema, scenario, cognitive model, folk model, etc. (Fillmore 1982, 111).

2. The project has been in development at The International Computer Science Institute in Berkeley since 1997.

1.2 Frame semantics lexical analysis

Frame semantics based lexical analysis comprises an analysis of the meaning of an LU, its lexical surroundings, phrases and grammatical constructions in which it appears in the corpus, the context in which it is used provided by corpus examples, as well as all the phrases in which the LU fulfills its full semantic potential. This approach consists of listing all LU arguments and adjuncts crucial to describing its meaning. Special attention is given to words that cannot be defined outside of the frames they are associated with. Those words are called frame-evoking words and are primarily verbs, but they also include nouns, adjectives and adverbs (Atkins, Fillmore, and Johnson 2003, 252).[3]

The basic units of a FrameNet analysis are frame and LU, a lexeme used in one of its senses (Fillmore et al. 2003, 297), (Ruppenhofer et al. 2016, 7).[4] In contrast to the standard lexicographic practice, which includes listing all the senses of a word in as much detail as possible, the LU in FrameNet is defined together with other LUs that belong to the same frame (Fillmore et al. 2003, 299)).[5] That is how, when we have defined the *Being_at_risk* frame, we can then define the nouns *risk*, *danger*, *safety*, *vulnerability*; adjectives *insecure*, *safe*, *secure*, *susceptible*, *vulnerable*, etc. with reference to the frame in question.

The process of describing a LU in FrameNet is defined in (Fillmore et al. 2003).[6] It begins with an informal description of the frame which a LU

3. The Frame semantic theory inspired us to point out the necessity of citing relevant constructions alongside the description of word meaning in the descriptive dictionaries of Serbian for all of the four most common frame-evoking word classes (nouns, adjectives, verbs and adverbs) (Марковић 2017, 34–41).

4. In Serbian lexicographic literature, as well as in syntax papers that explore the relationship between grammar and dictionaries, different terminology is used for what is referred to as lexical unit within FrameNet (e.g. in a university textbook of lexicology, that what is called a lexical unit refers to a lemma or a vocabulary entry, (Драгићевић 2007, 30), while Lj. Popović insists on shifting the focus to individual word senses and a lexeme used in one of its meanings is dubbed a sublexeme in his terminology (Поповић 2003, 202–203). In this paper, we decided to use the term *lexical unit* in order to stay within the framework's terminology.

5. Here we are referring to two approaches to describing lexical meaning, one that is *word-based* and the other *frame-based* (Atkins, Fillmore, and Johnson 2003, 254).

6. Although the process is described as an ordered sequence of steps, the authors still call for revising the data at any point and going back and correcting it if necessary (Fillmore et al. 2003, 299).

belongs to, a description of the situation or event represented by the frame and creating a list of words whose meaning would be described with reference to that frame (Fillmore et al. 2003, 299).[7] After that a target LU for which annotation is being done is chosen; that is typically one word but can be a multi-word unit or a phrase (Ruppenhofer et al. 2016, 21) and its use is looked into by extracting sentences, which contain it, from the corpus.

A lexicographer working in FrameNet compares his or her insight into the meaning of a target lexeme, based on corpus examples, to the meaning given in descriptive dictionaries.[8] Once he gets a clearer idea of its meaning, the lexicographer tries to describe the frame the LU belongs to more closely. After that, he writes the definition of the frame – a schematic description of an event which is central to a word, along with the names of participant roles called frame elements. The way in which frame elements are expressed in sentence examples of the target LU is lexicographically relevant (Fillmore et al. 2003, 304–305).

1.3 Frame elements

Frame elements have often been viewed as an extension of semantic roles (agent, experiencer, patient), but they are defined as *frame-specific*. This stems from a multitude of reasons, the most prominent being the ability to create a detailed definition of frame elements, which is not afforded when trying to fit the role into a predefined set (305).

First, the central elements of the frame (*core elements*) need to be identified.[9] Core elements are essential as they identify the frame as unique and set it apart from other frames. Alongside the core elements, there are *non-core*

7. That description entails:1) a schematic description of entity types or situation illustrated by the frame; 2) choosing descriptive labels for describing the frame; 3) drawing up a draft list of words that belong to the frame (if an LU belongs to a frame, it means that it can be subjected to the same analysis as other LUs in the frame) (Fillmore et al. 2003, 297).

8. Having analyzed the definition of the verb to risk in ten general-use dictionaries of English, Fillmore and Atkins concluded that even dictionaries of a similar size and purpose do not feature the basic meanings of the verb, which are part of basic vocabulary (Fillmore and S. Atkins 1994, 353).

9. There are some formal characteristics that help determine element centrality (e.g. core elements need to be expressed and so do those that have an interpretation even though they are not expressed (e.g. in the sentence *John arrived* the place where John arrived, the GOAL element, is not expressed but is still interpreted in the context (Ruppenhofer et al. 2016, 23–24).

elements that appear in all the frames in which an agent performs an action (they usually denote Place, Time, Manner, Instrument).[10] The situations where core elements are not linguistically expressed also occur, but they are still mandatory in the conceptual structure of the frame; this is called *null-instantiation* and is also annotated in the database (320). (Fillmore et al. 2003, 320). After the core and non-core elements are identified, we can move on to defining the frame itself.[11]

After analyzing the verb *to risk* in descriptive English dictionaries, Fillmore and Atkins discovered that not enough attention is given to its arguments (although they are very important for describing the word's meaning and essential in L2 English dictionaries) and that there are other sentence constituents that are completely overlooked in dictionaries, but need to be singled out and well-described in order to demonstrate correct verb usage. For example, an action performed by a person who is risking something (and can be syntactically expressed in multiple ways): She risked her life *trying to save a drowning child*; an objective someone has when putting themselves at risk: She risked her life *in order to save mine* (Fillmore and S. Atkins 1994, 362). An action by means of which someone takes a risk is one of the core elements of the frame, while the objective because of which they are taking it is non-core.

1.4 Frame-frame relations – FrameNet

After a frame and its elements are defined, a frame is connected to other frames. In that way frames, their elements and LUs belonging to them are placed in the semantic space (Ruppenhofer et al. 2016, 79) and make up a network. Creating frame-to-frame relations allows us to see and record semantic generalizations based on the type of participants, events, etc. A frame can be connected to frames it inherits from, has a perspective on, is perspectivized in, its subframes as well as the ones it uses. Frame-to-frame

10. The core/non-core distinction in the broadest terms corresponds to arguments and adjuncts in the traditional grammatical analysis (Fillmore et al. 2003, 310). Non-core elements cannot function as subject or object of the target verb and are often expressed by using an adverb or a prepositional phrase (319).

11. Ruppenhofer et al. (2016, 65) define other frame elements as well: elements that appear in subordinate clauses are non-core or extra-thematic e.g. TIME, MOTIVE. In addition to these, there are core-unexpressed elements that are considered core but do not have to be inherited by a child-frame (24–25). This paper does not get into detail about either of them.

relations are directed or asymmetrical: the more abstract and independent frame is called *Super_frame* and the more dependent and less abstract frame is called *Sub_frame* (Ruppenhofer et al. 2016, 79).

A list of frame-to-frame relations has been defined with the following ones being the most important (79–84):

- Within the *Inheritance relation* the Sub_frame is a more specific version of a more abstract parent frame. All the frame elements of the parent have a specified mapping with the frame elements of the child, while the child can have Sub_frames, FEs and semantic constraints specific only to itself (Fillmore et al. 2003, 311). For instance, the frame Run_risk Inherits from the frame Likelihood.
- The *Using relation* exists when a frame makes a general reference to the more abstract frame. An illustration of this would be the following frames: Wagering which uses the frame Run_risk; Speed which uses the frame Motion; Volubility which uses the frame Communication (Ruppenhofer et al. 2016, 83).
- *Perspective_on is* a relation similar to the broader relation of Using, but it puts greater constraints on the frames bound by it (82). In order for this relation to be possible, there need to be at least two perspectives for viewing a neutral frame. For instance, the frame *Risk_scenario* is a neutral frame, while the frames *Risky_situation*, *Being_at_risk* and *Run_risk* are all perspectivized; the situation is viewed from the perspective of one of the participants. The frames *Hiring* and *Get_a_job* are both perspectives on a neutral form of *Employment_start*, from employer and employee perspective.

After the definitions of the frames and their elements have been entered into the database, LUs can be added to the frames (in the case of *Being_at_risk*, the LU *risk* would be added). This is followed by the information on word class, meaning, formal composition (whether it is a single word or a multiword expression), after which instructions are given on how the corpus [12] can be searched in order to extract the concordances (subcorpus) that contain the exact lexeme we are looking for (in our case the noun risk) whose grammatical form points to the LU which belongs to the frame *Being_at_risk*. The aim is to weed out all the instances in which the searched keyword does not represent the LU that belongs to the frame which is being created. After the suitable searches for the desired LU have been specified, a number of

12. Fillmore et al. (2003, 304) use British National Corpus.

automated processes generate a subcorpus ready for annotation. This subcorpus is then cleaned of sentences that are too long or in any other way inadequate, and from those three to five sentences are chosen for each pattern with the aim of illustrating the variety of existing patterns rather than their statistical representativeness.

When the annotation is over, tools for analyzing the annotated sentences and the valence patterns instantiated within them are used. There are two types of reports in the form of dynamic web-pages (*LexUnit Report* and *Lexical Entry Report*) which are automatically generated after the annotation is finished and are available on the FrameNet website. The first report shows all the annotated sentences for an LU. Moreover, all the elements found in the current frame are listed (in a table of frame elements) and each element is color coded in the table, as well as in the annotated sentence. The second report gives an overview of the syntactic realizations of the frame elements and LU valence patterns in two tables (Fillmore et al. 2003, 326–328).

Since FrameNet also annotates frame elements (for frame-specific semantic roles) and their lexical realizations, terms like *valence group*, *valence pattern* and *valence description* are also important.[13] A frame element, together with its grammatical realization (unit type and its role in a sentence) constitutes a valence group, a set of valence groups used in a sentence makes up a valence pattern and the set of all valence patterns that a particular LU uses makes up a valence description (Atkins, Fillmore, and Johnson 2003, 255–257).

1.5 Different applications of FrameNet

FrameNet is available on the website. It can be searched and scrolled through online, but also downloaded and used locally. As the website states, it can be used for different purposes: as a dictionary for language learning (since it contains more than 13,000 LUs); as a valence dictionary; as a training dataset for semantic role labeling[14] which makes it a rich digital language resource (with over 200,000 manually annotated sentences linked to over 1,200 semantic frames).

13. The property of verbs to take arguments is called *valence*. Depending on the number of arguments they take, verbs can be: *monovalent* (when they require a subject), *divalent* (when they require a subject and an object), etc.

14. Subsection 1.6 will give an overview of some of the research done on the use of FrameNet and semantic role labeling programs for Croatian, Slovenian and Serbian.

FrameNet was conceived as a lexical database of English, which incorporates the databases subsequently developed for other languages (French, Chinese, Portuguese, German, Spanish, Japanese etc.) as part of various independent projects, applying the same formal structure and concepts. A project for aligning the data created for different languages has also been launched.

1.6 Previous research

In this section we will look into the research done in the field of semantic role labeling for Serbian and the languages related to it, as well as into the research devoted to the meaning of the noun risk and the verb to risk in discourse.

In the paper (Gantar et al. 2018) a model of semantic role labeling for Slovenian and Croatian was presented that they had developed as part of the international bilateral project *Semantic Role Labeling in Slovene and Croatian*. The objective was to develop a manually annotated corpus that would be used as a training dataset for supervised machine learning systems. An automatic semantic role labelling experiment, based on supervised machine learning is also described in the paper. The most frequent verbs, semantic roles and typical semantic-syntactic patterns of the most frequent verbs were presented for each of the corpora. The verb *to be* and the semantic role of patient were the most frequent in both corpora, while the second place went to the role of agent (95–96). In the paper, semantic roles were labeled in stable semantic-syntactic models (96–97), but the question of whether this is a valid method remains because semantic roles and frames are formed around a LU, a (verb) lexeme in one of its senses.

The paper Brač and Anić (2019) showcases a project aimed at developing a methodology for semantic-role labeling in a domain-specific language (in their case the domain of aviation) that could also be used in other fields. The authors of the paper examined whether it would be better to use more general semantic roles or verb-specific and frame-specific roles, typical of FrameNet. They came to the conclusion that too many specific semantic roles slow down the annotation process, but do not, in turn, contribute significantly to the improvement of terminology resources, although they noted that the list of broader semantic role labels needed to be slightly expanded (545).

The paper Wasserscheidt and Hrstić (2020) presents interesting research done for Serbian and Croatian (viewed as varieties of one language) on lexemes that both enter the general lexicon and form part of a certain professional domain (in this case legal terminology). It focused on whether or not

they take different meaning (evoke different frames) in Serbian and Croatian. The idea came from the authors noting a contradictory stance in the literature on frame semantics. Namely, Fillmore's works point to a difference in frames that individual speakers, social groups and cultures have, but later papers by other authors overlook this fact and treat frames as universal language-independent structures (88–89). The authors of the paper explored the meaning of the word *odredba* (section of a legal act) within the legal framework and the general lexicon (where it can be used as a synonym for a legal act as a whole) in both Serbian and Croatian corpus data. They used distributional analysis whose main tenet is that word meaning can be defined based on the context in which the word appears, and additionally applied the analysis on the context itself. Frame semantics theory was used to analyze the context (90). In view of the findings of these two distributional analyses, the authors concluded that there was no significant difference in the meaning of *odredba* in the corpora under examination and that the method of double clustering can be used in complex semantic analyses, which can then be represented through FrameNet structures (108).[15]

Although not directly related to our topic of FrameNet, we would still like to mention a paper that notes that *risk* has become a prominent topic in social science research with the research into the meaning of the word itself remaining vaguely defined (Hamilton, Adolphs, and Nerlich 2007, 164). Guided by this notion, the authors continue to analyze the meaning of the noun *risk* and the verb *to risk* using *the Cambridge and Nottingham Corpus of Discourse in English*, abbreviated as CANCODE. Analyzing the semantic tendencies of these lexemes and their semantic prosody, they conclude that the target lexemes are influenced by the context in which they appear (for example, there is a difference between their collocations and semantic prosody in a more intimate setting between family members and partners as opposed to student-professor exchanges).

2 A Couple of Instances from the Risk Domain

As cited above, at the end of Subsection 1.3, Fillmore and Atkins discuss the constraints on lexical analysis put by the traditional approaches to lexicography and the form of descriptive dictionaries (Fillmore and B. T. Atkins 1992, 100–101), (Fillmore and S. Atkins 1994, 350–363). After they juxta-

15. The analysis indicated that *odredba* is part of as much as 12 frames (Wasserscheidt and Hrstić 2020, 108).

posed the analyses done for the verb *to risk* and the noun *risk* in monolingual dictionaries and corpus data, they concluded that the dictionaries do not give a comprehensive enough description, with a lot of the meanings found through corpus search not even being mentioned. The finding was that printed dictionaries, with a linear approach to meaning, cannot represent a complex description needed to provide all the data of significance for the ways in which a word is used. This was the motivation for creating an online dictionary whose entries are frames rather than lexemes, as found in paper dictionaries, providing a notation better suited to such a complex system.

Conceived in such a manner, an online dictionary allows for representation of individual frame elements and their diverse syntactic realizations and therefore a full description of an element's valence (described in Subsection 1.4) as well as the relations between frames.

A visualization tool for viewing the relations between frames and their FEs (*FrameGrapher*)[16] makes it possible to choose the target frame and explore its relations to other frames. Figures 1–4 in this paper have been generated using this tool.

2.1 Frame Risky_situation (*Ризична_ситуација*)

The frame Risky_situation is shown below.[17] After giving a definition, we see illustrative examples in the form of sentences, as well as core and non-core elements of the frame. As mentioned above, all the FEs are color coded, with the same color that is used in the FE list appearing in the definition. The LUs evoking the frame Risky_situation are: *опасност.n* (*danger.n*), *опасан.a* (*dangerous.a*), *ризик.n* (*risk.n*), *рискантно.adv* (*riskily.adv*), *ризичан.a* (*risky.a*), *безбедан.a* (*safe.a*), *безбедно.adv* (*safely.adv*), *небезбедан.a/шкодљив.a* (*unsafe.a*), *претња.n* (*threat.n*). Frame-evoking LUs in the annotated example sentences are highlighted in black. A definition is given for each FE and followed by an example of its use.

16. FrameGrapher

17. For the purpose of this paper, we took original English frames and their elements, based on the data from English language corpora, and translated them into Serbian in order to illustrate the way of presenting data in FrameNet. It is our hope that we will soon get a chance to illustrate frames using Serbian corpus data.

Ризична_ситуација

Дефиниција:

Одређена Ситуација може (али не мора) да доведе до штетног догађаја који би задесио неку Вредност. Та Ситуација може бити неко стање, активност или неки кључни ентитет који мора бити схваћен као део неке шире Ситуације, која укључује и тај ентитет и Вредност. Иако је за разумевање овог оквира кључна идеја о штетном догађају, он не мора бити изражен као аргумент лексичких јединица у овом оквиру.

Да ли су климатске промене ОПАСНЕ по човечанство?
Купци се могу жалити на ШКОДЉИВЕ производе.
Највећа ОПАСНОСТ прети нашој инфраструктури.

Елементи оквира

Централни:

Вредност(Asset) [ass]	Нешто што се сматра вредним и пожељним, а постоји могућност да ће му штета бити нанета или да ће бити изгубљено. Мед није БЕЗБЕДАН за бебе.
Опасан ентитет (Dangerous entity) [dan ent]	Конкретан или апстрактан ентитет који може нанети штету Вредности или довести до њеног губитка. Том човеку је РИЗИК друго име!
Искључује: Ситуацију	
Ситуација (Situation) [sit]	Ситуација може довести до неког штетног догађаја. Већина људи се слаже да није БЕЗБЕДНО возити брже од 120 km/h.

Периферни:

Околности (Circumstances) []	Околности под којима је Вредност угрожена.
Степен (Degree) [deg]	Одредба која изражава одступање тренутног нивоа безбедности од очекиване вредности, узимајући у обзир Ситуацију и стање на које указује циљна ЈЛ. Терористи су наша највећа ПРЕТЊА.
Домен (Domain) [dom]	Домен у коме је Ситуација безбедна. Све наше сајтове *је* БЕЗБЕДНО користити у образовању.
Учесталост (Frequency) []	Колико често Вредност долази у *Ризичну_ситуацију*.
Место (Place) [pla]	Одређена локација на којој је Ситуација безбедна. Често се може закључити да карактеристике неке локације чине одређене Ситуације безбедним или небезбедним.
Време (Time) [tim]	Временски период током ког одређена Ситуација има прецизирани ниво сигурности.

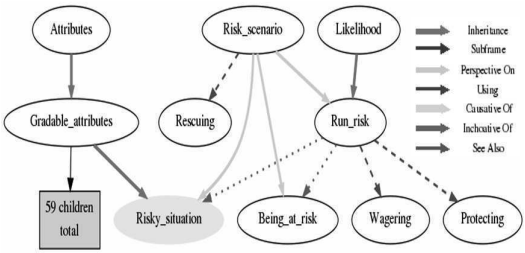


Figure 1. An illustration of the frame *Risky_ situation* and the related frames

2.2 Frame Being_at_risk (*Бити_угрожен*)

The LUs which evoke the frame Being_at_risk are: *опасност.n* (*danger.n*), *несигуран.a* (*insecure.a*), *ризик.n* (*risk.n*), *безбедан.a* (*safe.a*), *сигуран.a* (*secure.a*), *безбедност.n* (*safety.n*), *поуздан.a* (*reliable.a*), *рањивост.n* (*susceptibility.n*), *рањив.a* (*susceptible.a*). This frame contains the same FEs as the previous frame with the addition of Harmful_event (*Штетан_догађај*) and has the same color coding.

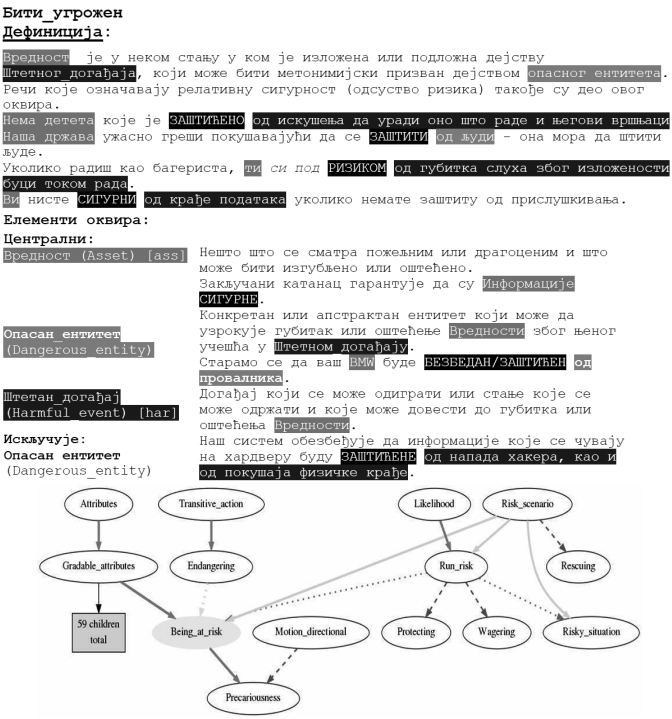


Figure 2. Semantic frame *Being_at_risk*

2.3 Frame Run_risk (*Изложити_се_ризику*)

The LUs evoking the frame Run_risk are: *угрожен.a* (*endangered.a*), *опасност.n* (*peril.n*), *ризик.n* (*risk.n*), *ризиковати.v* (*risk.v*), *угрозити.v*

(*endanger.v*). The definition, examples and FEs of the frame are given in Figure 3.

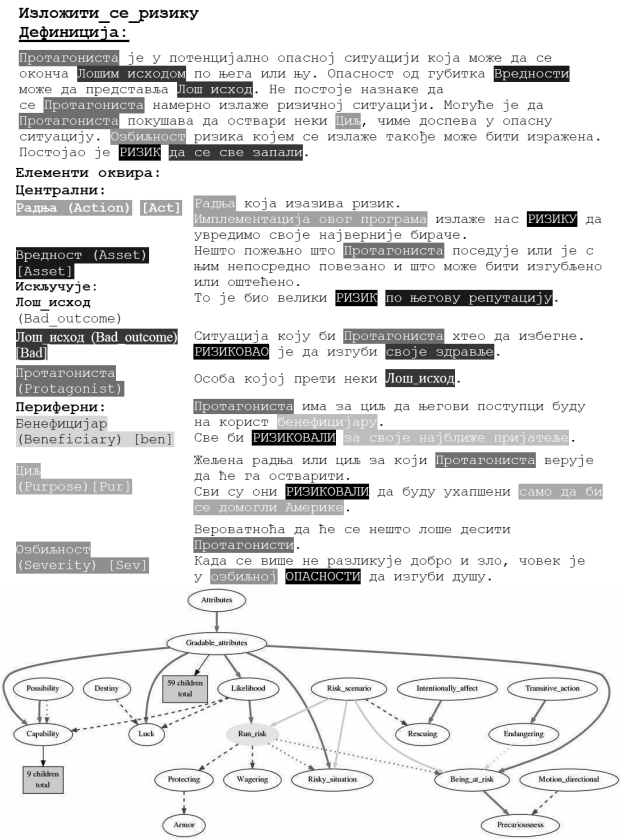


Figure 3. Semantic frame Run_risk

2.4 Frame Risk_scenario (*Сценарио_ризика*)

Figure 4 illustrates the relations between the frame Risk_scenario *(Сценарио_ризика*) and frames Run_risk (*Изложити се_ризику*) and Risky_situation (*Ризична_ситуација*) whose characteristics are shown in detail with their core (abbreviated as c) and non-core (abbreviated as *nc*)

elements listed. On the right-hand side there is a legend showing different types of frame-to-frame relations e.g. Inheritance, Perspective on, Using (as well as some of the relations we did not mention: Causative of, Subframe, etc.).

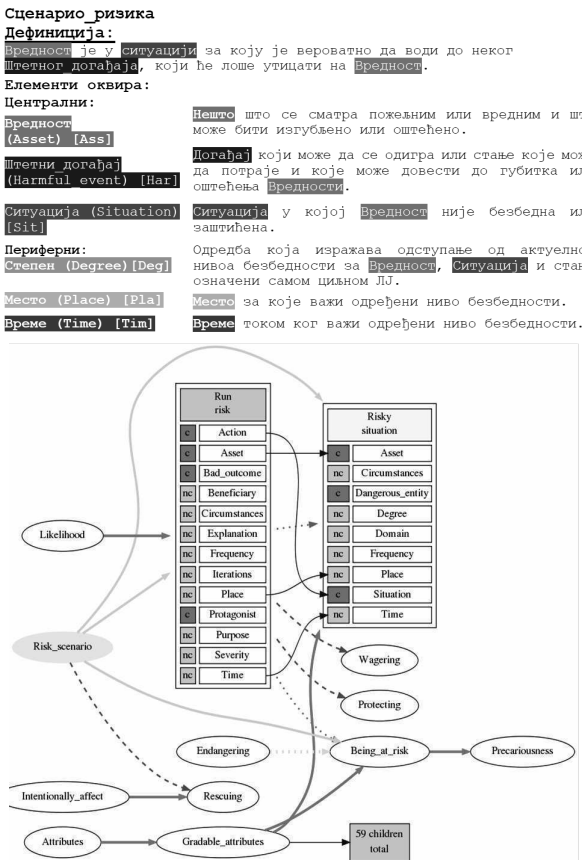


Figure 4. Semantic frame *Risk_ scenario* with a detailed view of two other related frames)

3 NLTK FrameNet Wrappers

NLTK (*Natural Language Toolkit*) is an easy-to-use natural language processing Python suite that accesses continually increasing number of corpora and lexical resources. NLTK offers different types of text processing, amongst which are: classification, tokenization, stemming, tagging, parsing and semantic reasoning. The NLTK system uses wrappers for other Python natural language processing and lexical resource libraries. One of the APIs available within NLTK is FrameNet and the accompanying program library designed for searching this resource, as well as for extracting information from it.

As mentioned in the Introduction (Section 1.1 of this paper), a frame is a conceptual structure describing a type of situation, entity or relation together with its participants. The structure of FrameNet within the NLTK framework is comprised of a collection of XML (*Extensible Markup Language*) files catalogued as: *frame*, *fulltext*, *lu*, *miscXML*, which are accessed through the library's commands or can be directly searched and visualized by means of XML files using XSL (*eXtensible Stylesheet Language*) transformations: *frameIndex*, *luIndex*, *fulltextIndex*. In this section, we will show the use of the FrameNet wrapper.

The function `frames()` lists all the frames contained in FrameNet. The following lines of code illustrate the initialization of working with FrameNet and return the information that the FrameNet version available in NLTK contains 1221 frames.

```
from nltk.corpus import framenet as fn
len(fn.frames())
```

In order to find all frames that contain the word *risk*, we use the command:

```
fn.frames(r'risk')
```

which outputs the following information:

[<frame ID=1560 name=Being_at_risk>,
<frame ID=378 name=Run_risk>].

Since the query is case-sensitive, we need to do a second search in order to find all the instances in which *risk* appears:

```
fn.frames(r'Risk')
```

which outputs a different result:

[<frame ID=1763 name=Risk_scenario>,
<frame ID=1762 name=Risky_situation>]

If the function `frame()` is given a regular expression '`(?i)risk`' as an argument, we get a combined list of the two, containing all four frames (sections 2.1–2.4), whose names correspond to the given pattern because '`(?i)`'expresses that the case of the letter is irrelevant.

The details of a frame can be listed through the command `frame()`, which is given the number of the frame as an argument, for instance `f=fn.frame(1762)`, returns all the data of the frame Risky_situation.[18]

Individual components of the frame can be accessed separately through the commands like: `f.name` giving the name of the frame, `f.definition` giving its definition, `f.FE` listing the elements of the frame, `f.lexUnit` giving frame LUs, `f.frameRelations` giving frame relations, as shown in the following example:

```
f = fn.frame('Risky_situation')
print(sorted([e for e in f.FE]))
print([r for r in f.frameRelations])
```

that outputs:

['Asset', 'Circumstances', 'Dangerous_entity', 'Degree', 'Domain',
'Frequency', 'Place', 'Situation', 'Time']
[<Parent=Gradable_attributes – Inheritance →
Child=Risky_situation>,
<MainEntry=Run_risk – See_also →
ReferringEntry=Risky_situation>,
<Source=Run_risk – ReFraming_Mapping →
Target=Risky_situation>,
<Neutral=Risk_scenario – Perspective_on →
Perspectivized=Risky_situation>]

4 Lexical Analysis of the Word *Risk* in a Mining-related Corpus

The development of a monolingual corpus in the domain of mining started as part of a mining project documentation management project using language

18. Data for the frame *Risky_situation*

technologies (Tomašević et al. 2018, 996). Back then, the corpus contained texts from the domain of mining and similar research areas with a total of 172 documents (in Serbian) and 2.7 million words in the first iteration (997). In the course of further research, 63 documents have been added (Kitanović 2021). The current version contains 4.1 million words. It comprises project documentation (26%), legislation (11%), doctoral dissertations (31%), textbooks and other mining literature (32%) (Kitanović et al. 2021, 8).


Concordances – Profil 1 – Microsoft Edge

https://leximirka.jerteh.rs/CQP/Noske

A(N)

na najmanju moguću meru , odnosno otklanjanje	profesionalnih rizika	. Strategija teži da se u ovom periodu broj
na najmanju moguću meru , odnosno otklanjanje	profesionalnih rizika	. Strategija teži da se u ovom periodu broj
inspektora rada sa novim tehnologijama i	novim rizicima	, savremenim pristupima i praksama u oblasti
i zdravlja na radu uzimajući u obzir	posebne rizike	koji se pojavljuju u određenim delatnostima . o
uzajamna povezanost , što samo još povećava	potencijalne rizike	za ukupnu realizaciju procesa rekultivacije .
velikih količina otpada po konkurentnoj ceni a	niskom riziku	po životnu sredinu ; 2 . ekonomski isplativ
, potencijalno moguće ozbiljnije povrede ,	mali rizik	fatalnog kraja , gubici radnog vremena Nizak
i normalna komunikacija Neophodna prva pomoć ,	mali rizik	od ozbiljnih povreda Zanemarujući Nemerljivi
6 , jer osim snabdevanja gasom , kod nje postoji i	izvesni rizik	od povraćaja investicije , što bi moglo
odgovora često veoma zahtevan , složen , i sa	prisutnim rizicima	. Konačno rešenje , kako smo već istakli u
i sl . • Prisustvo konfliktnih situacija ,	povišenih rizika	i nepovoljnih događaja , npr . interakcija
sistema zaštite na radu : 1) radnim mestima sa	povećanim rizikom	; 2) zaposlenima raspoređenim na radna mesta
; 2) zaposlenima raspoređenim na radna mesta sa	povećanim rizikom	i lekarskim pregledima zaposlenih
može da ima previd pojedinih opasnosti . Psiho -	socijalni rizici	se obično previde , kao i rizici u vezi sa
, a takođe je ostvaren napredak i u proceni	profesionalnih rizika	i sistematizaciji profesionalnih bolesti .
ili smanjenja rizika . Radno mesto sa	povećanim rizikom	jeste radno mesto utvrđeno aktom o proceni
je da se nekontrolisane opasnosti prevedu u	kontrolisani rizik	i da se na taj način bolje zaštite zaposleni i
identifikovanju i kontrole zdravstvenih i	sigurnosnih rizika	organizacije i eliminisanju ili smanjivanju
organizacije i eliminisanju ili smanjivanju	potencijalnog rizika	od nezgoda na prihvatljiv nivo , poštujući pri
najvišeg mogućeg nivoa bezbednosti i	minimalnog rizika	moraju se dokumentovati uključujući i zapise

Figure 5. Concordances for adjective-noun pattern containing the noun *ризик*

The results of a CQL[19] (*Corpus Query Language*) query are analyzed for: frequency lists, collocations, concordances with a narrower and broader context. Figure 5 shows the concordances extracted from the Leximirka[20] digital dictionary management web app (Stanković et al. 2018) of the adjective-noun pattern containing the noun ризик (risk), while in Figure 6 there is a histogram of frequencies for different inflected forms of the same pattern taken from a mining corpus, available on the open-source platform *NoSketch Engine* (Kilgarriff et al. 2004).[21]. The version on the local servers is maintained

19. Corpus Querying
20. LeXimirka
21. NoSketch at JeRTeh, NoSketch Engine

by members of the JeRTeh Society for Language Resources and Technologies.[22] A *Treegger* model for Serbian was trained for tagging (Krstev and Vitas 2005; Utvic 2011), (Stanković et al. 2020, 3957) using a manually annotated corpus of Serbian morphological dictionaries (Krstev 2008).

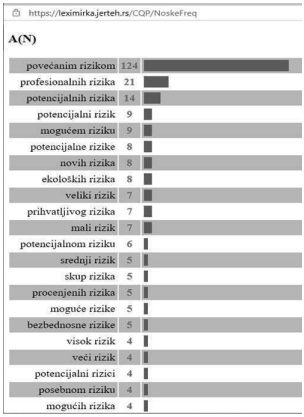


Figure 6. A histogram of frequencies for different inflectional forms of the noun *ризик*

The mining corpus is published in *Sketch Engine*[23] too (Kilgarriff et al. 2014), a platform that provides the option of different types of searches. For instance, we can extract concordances for a target lemma or multi-word expression, collocates of a lemma, related-word thesaurus, *Word Sketch* or *Word Sketch Difference* for two related words. The word sketch approach, developed by Kilgarriff et al. (2004), helps build FrameNet and similar resources and speeds up the process of sense disambiguation of polysemous words (Baker 2012, 274).

Word sketch gives a quick overview of the behavior of the target lexeme by gathering information from thousands or millions of examples of its use and summarizes collocates by category, with links to individual examples. Figure 7 illustrates the word sketch for the noun *ризик* – one look at the page gives a clear idea of the word's use. The first column shows prepositional phrases (in Serbian linguistic terminology referred to

22. JeRTeh
23. Sketch Engine

as *предлошко-падежна конструкција*):[24] *risk of/with/in/on/for...* (*ризик од/са/у/по/на/за...*), and if we clicked on "..." we would get the concordances for each individual phrase. The second column features the modifiers of the word, in this case passive participles of verbs: *increased* (*повећан*)/ *identified* (*идентификован*)/ *assessed* (*процењен*)/... *risk* (*ризик*) or adjectives: *potential* (*потенцијалан*)/ *political* (*политички*)/ *unacceptable* (*неприхватљив*)/ ... *risk* (*ризик*). The third column contains the verbs with which *ризик* appears as the subject e.g. *to decrease* (*смањити*)/ *to arise* (*настајати*)/ *to exist* (постојати)/.... What follows are the expressions in which *ризик* appears as an object: to *decrease* (*смањити*)/ to assess (*проценити*)/ ... *risk* (*ризик*).

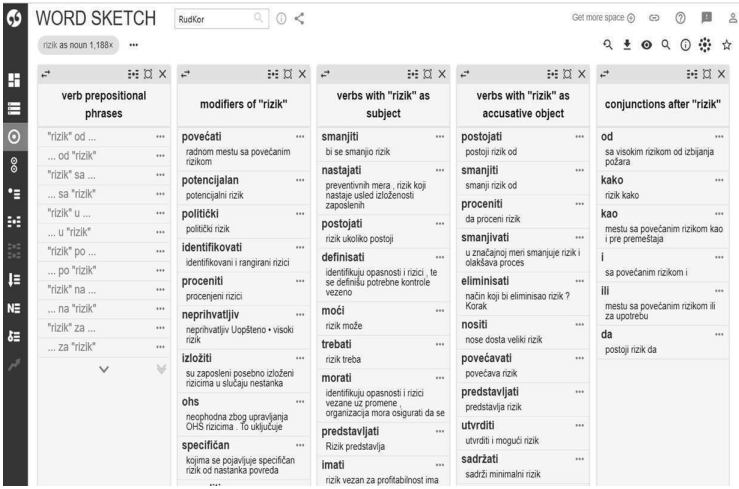


Figure 7. Sketch of the word *ризик* on *Sketch engine*

Figure 8 shows a dynamic diagram of the collocations. It is clear that most of the collocations are prepositional phrases. On the right-hand side of the picture there is the settings option allowing to choose which patterns are to be shown and the minimal frequency requirement that collocations have to meet in order to be included in the diagram.

24. It should be mentioned that the tools and automatic detection are not that well-suited for Serbian but are nevertheless valuable. Namely, mistakes are found that need to manually be corrected.

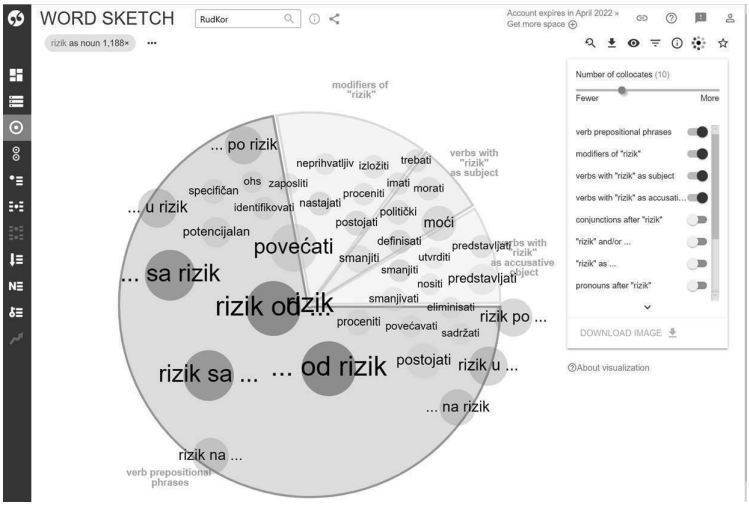


Figure 8. Illustration of collocations of the noun *ризик* in *Sketch engine*

Collocations research is very important (for example, in lexicography, it is important to list the most frequent collocates of a LU; collocations are crucial not only in language learning, but also in different natural language processing tasks). Using the word sketch and the collocation *risk of* (*ризик од*) as a starting point, a detailed view of the concordances can be shown (Figure 9).

The sketch gives a quick search with preset rules, but a custom search can be executed with CQL queries. If we wanted to see where the risk was coming from we would get an answer with the following query `[lemma="ризик"][tag="N"]`. The query `[tag="A"][lemma="ризик"]` would give an answer to the question what type of risk it is; or, if we allowed the result to contain examples in which no more than 5 words divide our target word and the verb we would write the following query: `[tag="V"][word!=""]0,5[lemma="ризик"]`.

The frequencies of collocations can be both listed and presented visually with bars as shown in the picture below. Figure 10 shows the frequencies of the collocations containing the noun ризик.

Figure 11 illustrates Word Sketch Difference (an extension of Word Sketch). It generates a sketch of two target words and compares them, which allows for a clear overview of the differences in their use. This option is particularly valuable for similar meaning words, for antonyms and

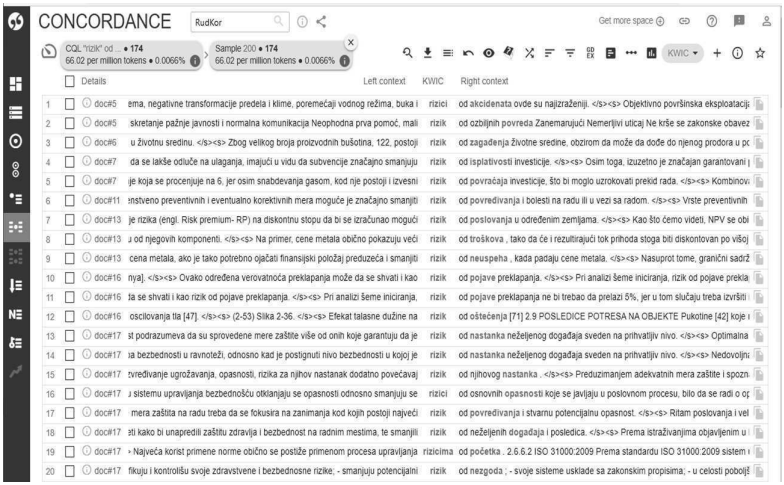


Figure 9. Concordances for *ризик од* in *Sketch engine*

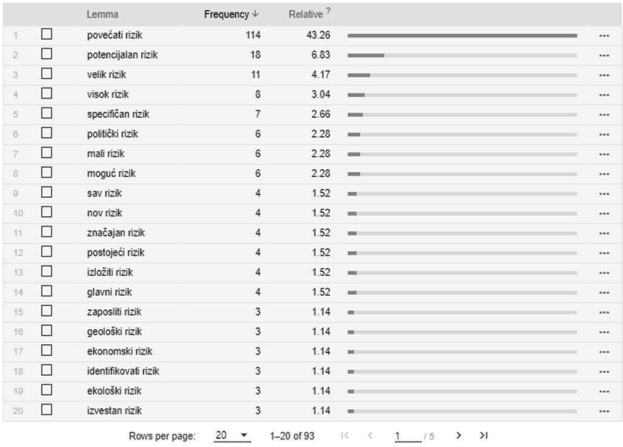


Figure 10. Collocation frequencies for the noun *ризик* in *Sketch engine*

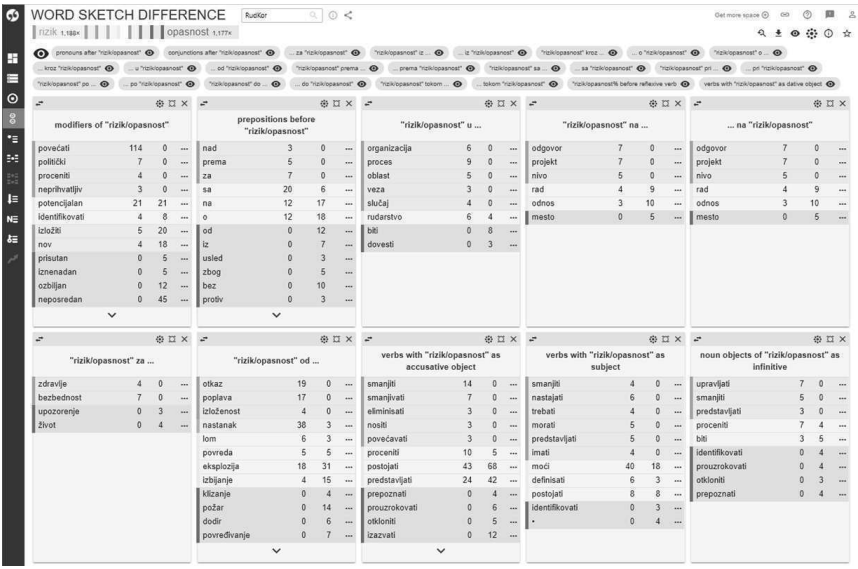


Figure 11. Word Sketch Difference of the words *ризик* and *опасност*

words from the same semantic field. It is shown in Figure 10 that the noun *risk* (*ризик*) has as its most frequent collocates: to *increase* (*повећати*), *political* (*политички*), *to assess* (*проценити*), *acceptable* (*прихватљив*), while the most frequent ones of the noun *danger* (*опасност*) are adjectives *непосредан* (*immediate*), *озбиљан* (*serious*), and *изненадан* (*sudden*).

The automatically generated thesaurus for the target word finds synonyms or words that fall in the same category (same semantic field) and lists them in a table with links to the sketches of individual words, concordances, word sketch differences and thesauruses. Figure 12 shows an illustration of the thesaurus which contains automatically retrieved words from the same semantic field as the target word *risk* (*ризик*), on the left-hand side in the form of a bubble graph and on the right-hand side as a word cloud. The thesaurus word list is created based on the context in which the searched word appears within a chosen corpus, relying on the distributional semantics theory, which, in short, postulates that words that appear in the same context have a similar meaning. In order to determine synonyms, word sketches for all words belonging to the same part of speech are compared and the words

that share the most collocates are paired as similar. The grade[25] given to each of the synonym points to the number of shared collocates.

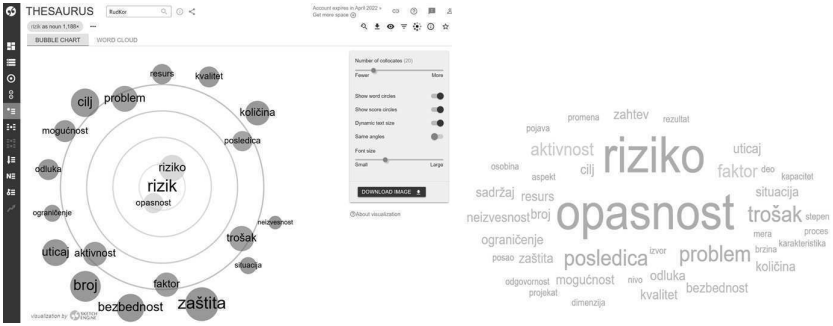


Figure 12. Illustration of the word's *ризик* thesaurus

5 Conclusion

This paper illustrates the results of preliminary research exploring the possibility of application of the frame semantics theory and the principles used in building the FrameNet semantic network using the examples from the risk domain adapted to Serbian. We also show the inner workings of the NLTK suite usable for many different language resources, as well as the Sketch Engine corpus analysis tool.

We have shown that FrameNet offers a detailed and structured mapping, which can then be used in different ways for language processing, especially in text extraction and organizing, as well as in an effort to make human-computer interaction more natural in applications like chatbots. A chatbot needs to be able to recognize different lexical units that evoke the same event or refer to the same entity in order to successfully recognize intent.

It is of great importance that the English FrameNet can be filled with entries from other languages e.g. Serbian (keeping frame information which is shared and adding language-specific material) therefore making it applicable to multilingual resources.

25. Статистичке формуле које се корите у алату *Sketch engine*: statistics/formulae

The research presented above only hints at the possibility of adapting FrameNet to Serbian and aligning that network with the FrameNet data in other languages. Future research intends to align the use of Serbian WordNet and Serbian FrameNet, joining them together. While working toward this aim, we will be following the recommendations given by Tonelli and Pighin (2009).

This research is also aimed at encouraging the growth of Serbian corpus lexicography efforts and modernization of the description of the grammar and lexicography of this language. A good step forward in the modernization process would be case studies that compare polysemous lexeme entries from the SASA (Serbian Academy of Sciences and Arts) dictionary to their description using the frame semantics analysis.

The possibilities for future research on this topic are vast. The implementation of frame semantics theory and methodology used in FrameNet, as well as the discussed tools, will pose a challenge for Serbian. Based on this paper, we speculate that it will be very challenging to use the concept of null instantiations to explore transitive verb complements which do not have to be overtly expressed and are, therefore, implicit (e.g. verbs *to cook*, *to write*, etc.), as well as to look into the ways in which descriptive dictionaries of Serbian deal with such phenomena. We also believe it would be useful to introduce this notion (three types of null-instantiation are defined in FrameNet) into Serbian grammar.

Acknowledgment

This paper was supported by the Ministry of Education, Science and Technological Development, of the Republic of Serbia, in accordance with Contract No. 451-03-9/2021-14, which was entered into with the SASA Institute for Serbian Language. Sketch Engine access is provided by the ELEXIS project funded by the European Union Horizon 2020 research and innovation program under grant number 731015.

References

Atkins, Beryl T. S. 1994. "Analyzing the verbs of seeing: a frame semantics approach to corpus lexicography." In *Annual Meeting of the Berkeley Linguistics Society,* 20:42–56. 1.

Atkins, Sue, Charles J Fillmore, and Christopher R Johnson. 2003. "Lexicographic relevance: Selecting information from corpus evidence." *International Journal of lexicography* 16 (3): 251–280.

Baker, Collin F. 2012. "FrameNet, current collaborations and future goals." *Language Resources and Evaluation* 46 (2): 269–286.

Boas, Hans C., and Ryan Dux. 2017. "From the past into the present: From case frames to semantic frames." *Linguistics Vanguard* 3 (1): 20160003. https://doi.org/doi:10.1515/lingvan-2016-0003.

Brač, Ivana, and Ana Ostroški Anić. 2019. "From concept definitions to semantic role labeling in specialized knowledge resources." In *Proceedings of the 13th International Conference of the Asian Association for Lexicography,* 604–611.

Fillmore, Charles J. 1976. "Frame semantics and the nature of language." In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech,* 280:20–32. 1. New York.

Fillmore, Charles J. 1982. "Frame semantics." *Linguistic society of Korea (ed.), Linguistics in the morning calm,* 111–137.

Fillmore, Charles J, and Beryl T Atkins. 1992. "Toward a frame-based lexicon: The semantics of RISK and its neighbors." *Frames, fields and contrasts: New essays in semantic and lexical organization* 75:102.

Fillmore, Charles J, and Sue Atkins. 1994. "Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography." In *Computational Approaches to the Lexicon,* edited by Sue Atkins and Antonio Zampolli, 349–393. Oxford: OUP.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Gantar, Polona, Kristina Štrkalj Despot, Simon Krek, and Nikola Ljubešic. 2018. "Towards semantic role labeling in Slovene and Croatian." In *Proceedings Conference on Language Technologies and Digital Humanities in Ljubljana,* 93–98.

Gildea, Daniel, and Daniel Jurafsky. 2002. "Automatic labeling of semantic roles." *Computational linguistics* 28 (3): 245–288.

Hamilton, Craig, Svenja Adolphs, and Brigitte Nerlich. 2007. "The meanings of 'risk': A view from corpus linguistics." *Discourse & Society* 18 (2): 163–181.

Jurafsky, Dan, and James H Martin. 2020. "Semantic Role Labeling and Argument Structure." Chap. 19 in *Speech and Language Processing,* 3rd ed. December 30, 2020 draft.

Kilgarriff, Adam, Vıt Baisa, Jan Bušta, Miloš Jakubıček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychl, and Vıt Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. "Itri-04-08 the sketch engine." *Information Technology* 105 (116).

Kitanović, Olivera. 2021. "Ontološki model upravljanja rizikom u rudarstvu." PhD diss., Univerzitet u Beogradu, Rudarsko-geološki fakultet. https://uvidok.rcub.bg.ac.rs/bitstream/handle/123456789/4305/Doktorat.pdf?sequence=1.

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A Data Driven Approach for Raw Material Terminology." *Applied Sciences* 11 (7): 2892.

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

Krstev, Cvetana, and Duško Vitas. 2005. "Corpus and Lexicon-Mutual Incompletness." In *Proceedings of the Corpus Linguistics Conference,* 14:17.

Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H Martin, and Dan Jurafsky. 2005. "Semantic role labeling using different syntactic views." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),* 581–588.

Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice.* Revised November 1, 2016. https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of The 12th LREC – Language Resources and Evaluation Conference,* 3954–3962.

Tomašević, Aleksandra, Ranka Stanković, Miloš Utvić, Ivan Obradović, and Božo Kolonja. 2018. "Managing mining project documentation using human language technology." *The Electronic Library,* https://doi.org/10.1108/EL-11-2017-0239.

Tonelli, Sara, and Daniele Pighin. 2009. "New features for framenet-wordnet mapping." In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009),* 219–227.

Utvic, Milos. 2011. "Annotating the Corpus of Contemporary Serbian." *INFOtheca* 12, no. 2 (December): 36a–47a.

Wasserscheidt, Philipp, and Andrea Hrstić. 2020. "Legal Variation? A Frame Analysis of Croatian and Serbian in the Domain of Law." *Mediterranean Language Review* 27:87–112.

Драгићевић, Рајна. 2007. *Лексикологија српског језика.* Београд: Завод за уџбенике.

Марковић, Александра. 2017. "Однос граматике и речника – граматика инхерентна описним речницима српског језика." *Наш језик* XLVIII (1-2): 27–43.

Поповић, Љубомир. 2003. "Интегрални речнички модели и њихов значај за лингвистички опис и анализу корпуса." *Научни састанак слависта у Вукове дане* 31 (1): 201–220.

On *Corpus of English-Studies Students (KorSAng)* and Possibilities of Its Software Exploitation

UDC 81'322.2
DOI 10.18485/infotheca.2021.21.1.2

ABSTRACT: Corpus linguistics in Bosnia and Herzegovina and Republic of Serbia is not used enough. The reasons for this are many: the lack of language corpora, the fear of computer methods, and the still present traditional approach to data processing that is qualitative or does not go beyond descriptive statistics. We will often hear arguments against the computer method, such as that a large number of examples that are the result of a search query can impair the quality of the analysis and be misleading. However, the development of technology has also influenced the development of information literacy in all human activities, including the academic community. In this paper, we will try to explain how several departments of English studies in Bosnia and Herzegovina and Serbia in cooperation with The Society for Language Resources and Technologies – *JeRTeh* in Belgrade offered a solution to the student corpus, describe the process of collecting corpus until its final form and demonstrate what search options the corpus offers to its end users.
KEYWORDS: corpus linguistics, learner corpus, parallel corpora, corpus annotation, Corpus of English-studies students.

PAPER SUBMITTED: 16 August 2021
PAPER ACCEPTED: 2 September 2021

Minja S. Radonja
minja.radonja@ff.ues.rs.ba

Srđan R. Šućur
srdjan.sucur@ff.ues.rs.ba

University of East Sarajevo
Faculty of Philosophy Pale
Sarajevo
Bosnia and Herzegovina

1 Introduction

"Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular

SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance" ((Granger 2002, 7), as cited in (Марковић 2019, 13)). *Corpus of English-studies students* (Sr. *KorSAng*) is a learner corpora of Serbian students of English studies from 4 universities – The University of East Sarajevo, The University of Banja Luka, The University of Belgrade and the University of Novi Sad. The corpus is the product of two important national projects: *Phraseological competence of Serbian speakers of English through the prism of contrastive analysis of interlanguage*, 19/6-020/961-46/18, which was realized in the period from December, 31, 2018, and *Scientific potentials of annotated student corpora in applied linguistics*, 19.032/961-135/19, realized in the period from December, 31, 2019. The KorSAng corpus is inspired by the first student corpus of the English language of Serbophone speakers – *ICLE-SE*, which was formed in the period from September 2015 to September 2016 and consists of argumentative essays of Serbophone English-studies students written in English. Previous research results based on the *ICLE-SE* corpus are presented in scientific papers (Марковић 2017, 2018, 2019, 2020; Radić-Bojanić 2019; Radonja 2019; Šućur 2019; Spajić and Suknović 2019; Tomović and Stefanović 2019). See more about the *ICLE-SE* corpus and the methods of its search in (Шуњур 2020).

1.1 Corpus structure

KorSAng includes 327 texts and has a total of about 140,000 words. It consists of three subcorpora:

- Subcorpus of English-studies students' translations from English (KorPSAng1), which consists of translations of literary texts (3 texts) and newspaper articles (3 texts) from English into Serbian. This part of the corpus has 127 translations and 46,785 words. This practically means that each translation can be paired with the original text which is in this subcorpus in English, which is together called bitext.[1]
- Subcorpus of English-studies students' translations into English (KorPSAng2), where the initial texts are from the genre of literature

1. "[B]itext is a text and its translation, i.e. translations, presented in such a way that an explicit connection is established between the elements of their logical statement, for example, at the level of paragraphs or sentences" ((Vitas 2010, 273), as cited in (Андоновски 2019, 17).

(1 text) and the newspaper genre (5 texts). The final number of translations of this subcorpus is 130, and the number of words is 50,128. As the previous subcorpus, *KorPSAng2* is also a parallel corpus.
– The subcorpus of argumentative essays[2] of English-studies students written in Serbian as a mother tongue (KorSSAng) includes 70 essays and 40,012 words.

The first two components are parallel student corpora, while the third component is designed as a reference corpus to the ICLE-SE corpus. "Parallel corpora are understood to be corpora that contain one or more original texts and their translations into one or more languages" ((Töny 2016, 11), as cited in (Андоновски 2019, 16)). On the importance and application of parallel corpora see more details in (Андоновски 2019, 2021; Ристовић 2012, 2016).

Next, in the second section of this paper, we will describe the stages of corpus preparation, from collecting and selecting texts, through processing and parallelization of texts, to creating metadata and forming the corpus. The third section shows how the corpus can be searched. Finally, in conclusion, we summarize the results so far and present further plans.

2 *KorSAng* corpus preparation

A total of 194 English-studies students participated in the formation of the corpus. *KorSSAng* subcorpus was formed in the period from 2016 to 2019, and *KorPSAng1* and *KorPSAng2* from 2017 to 2019. Of these three components, the most demanding one to collect was a corpus of essays in the Serbian language, due to students' reluctance to participate in essay writing in their mother tongue, so that the formation of this part of the corpus took the longest. The project of creating a parallel *Corpus of English-studies students* consisted of several stages:

– preparatory phase,
– collecting texts,
– text processing,
– parallelization,
– keeping records of metadata,
– publishing texts and metadata on the platform.

2. "An argumentative text (discussion) is a text in which the author starts from some doubt or dilemma, and then, by stating the arguments, a solution is reached. The author of such a text strives to convince the interlocutor or the reader of the correctness of his opinion with the evidence" (CEO 2015, 79).

2.1 Preparatory phase

In the preparatory phase, several key issues related to the project were defined: main and secondary objectives of the project; determining the scope of the project; specifying cooperation and distribution of tasks among team members from the University of East Sarajevo, the University of Banja Luka, the University of Belgrade and the University of Novi Sad; specifying cooperation with team members from JeRTeh;[3] determining the time frame for the execution of individual stages. Jelena Marković, the project manager, was in charge of establishing contacts with the team members.

2.2 Collection and selection of texts

In the next phase, the team members, together with the project manager, selected several texts for translation from English and into English and made a preliminary list of the essay topics in Serbian. During the text selection, care was taken to ensure that the texts were from the literary and newspaper genres. Newspaper articles cover various topics, such as politics, economics, language, health, social networks, and show business. In the end, 6 texts in English, 6 texts in Serbian and 24 essay topics in Serbian were selected. Then ensued the collection of translations and essays, which was the most time-consuming task. It consisted of organizing the conditions for translating texts and writing the essays by the students. Team members organized this task at their faculties. The translation was limited in time, with the exception of a few works, while the writing of the essay was unlimited in at least a third of the works. In addition to the given written assignments, students were required to fill in the accompanying document profile of the participants in order to give consent for their translations and essays to be used for research purposes, while the papers themselves are anonymous and recorded under a code.

2.3 Processing and parallelization of texts

Work tasks related to text processing and parallelization were performed by research assistants, under the leadership of Jelena Marković and the team from JeRTeh. The assistants received online training in the use of

3. JeRTeh

parallelization tools to obtain texts in TMX (Translation Memory eXchange) documents.[4] They received instruction in Notepad++ XML editor,[5] Unitex/GramLab,[6] Unitex module for Serbian (Krstev 2008) and ACIDE (Aligned Corpora Integrated Development Environment) (Utvić, Stanković, and Obradović 2008), which enabled adequate text processing by prescribed rules. In Notepad++, the source and translated texts were prepared, which involved pairing paragraphs of two documents, so that each paragraph of the source text corresponded to a paragraph of the translation. The files prepared in this way were further processed in the *Unitex Visual IDE*, which enables the segmentation of paragraphs into sentences, according to the language in which the text is written. After segmentation of the original and translated text, the files were prepared for parallelization, i.e. "the process of establishing links between the appropriate variants of translation units, i.e. the formation of a set of translation units" (Андоновски 2019, 17). Parallelization was enabled using the *ACIDE* application developed by the Language Technology Group of the University of Belgrade (more on the *ACIDE* integrated development environment for parallelized corpora can be found in (Utvić, Stanković, and Obradović 2008). This application enables automatic parallelization with the possibility of checking and correcting errors, which helped the team members who had this task to successfully prepare texts for the parallel corpus. In the texts prepared for the parallel corpus within this project, we came across examples that some segments were missing in the translated text, or one segment in the original text corresponded to several segments in the translated text, which can occur in translation. The training itself was a challenge to a team of linguists who had not encountered this type of work tasks before, and the skills acquired during the creation of the parallel corpus gave a new perspective on corpus linguistics. The final result after word processing in *ACIDE* was a *TMX* document that was then entered into the corpus.

4. TMX is an ISO standard (ISO24616 2012) for the storage of so-called translation memories and their exchange between different software translation tools, as well as between different companies that deal with the maintenance of translation memories (TMX 2005). Translation memories are collections of determinants in which the text of the source language is associated with the equivalent translation of the text of the target language, i.e. the produced TMX document is composed of the obtained translation units (Андоновски 2019, 22).

5. Notepad++

6. Unitex/Gramlab, Cross-platform Corpus Processing Suite

2.4 Keeping records of metadata and corpus formation

In parallel with the preparation of texts for the corpus, the researchers kept records of metadata containing information on variables such as data about the participant (age, gender, mother tongue, education data, years of learning English, knowledge of other foreign languages, stay in English speaking country) and variables concerning the context of language production, i.e. data on the context in which the students performed the tasks (whether there was a time limit for the tasks of writing the essay or translation, use of dictionaries and manuals, whether the task was an exam obligation or written outside the exam, what is the genre, etc.). Metadata tables were made separately for translations and essays in separate Excel documents. Since the papers entering the corpus were anonymous, it was necessary to assign to each document the appropriate metadata registered under the same code. Team members from JeRTeh prepared the documents for publication: assigned metadata to each document, did part-of-speech annotation and lemmatization (for more details on the resources that made this possible, see (Stanković et al. 2020) Stanković et al. 2020).

3 Search of the *KorSAng* corpus

We briefly described the process of collecting the corpus from the preparatory phase to its final electronic form and the challenges we encountered along the way. *KorSAng* contains 136,925 words: *KorPSAng1* (46,785 words, 127 translations into Serbian), *KorPSAng2* (50,128 words, 130 translations into English), *KorSSAng* (40,012 words, 70 essays in Serbian). One of the major challenges was to motivate students to write essays in their mother tongue, which raises doubts about the decline in the competence of writing in their mother tongue (Šućur 2020, 144). We will also mention that the essay topics limit the search possibilities. Of the 24 topics offered, more than 40% of the essays were written on four topics: *Family or work: what is more important* (Sr. *Породица или посао: шта је важније*), *How music affects life* (Sr. *Како музика утиче на живот*), *How I imagine a good parent* (Sr. *Како замишљам доброг родитеља*), and *From marriage to divorce today* (Sr. *Од склапања брака до развода данас*). Therefore, we expect that the corpus of the essays offers a rich vocabulary that belongs to the semantic field of family relations, but poorer when it comes to some other offered topics. Next, after a few introductory remarks about the *Sketch Engine* platform, we shall first demonstrate how it can be used to perform simple and advanced searches

of the *KorSSAng* corpus, and then how it can be used to perform parallel corpora searches of the *KorPSAng1_ en* and the *KorPSAng1_ sr* corpus (the former consists of the source texts in English, and the latter consists of the Serbian translations of the source texts) corpus, and the *KorPSAng2_ sr* and the *KorPSAng2_ en* corpus (the former consists of the source texts in Serbian, and the latter consists of the English translations of the source texts) respectively.

3.1 *Sketch Engine*

The Sketch Engine is an online platform that comprises a series of robust electronic corpora query tools. This platform is the result of a collaboration between the British lexicographer and corpus linguist Adam Kilgarriff, and the Czech computer scientist Pavel Rychlý. Its commercial version has been available since 2004 (Kunilovskaya and Koviazina 2017, 503). As Kilgarriff et al. (2014, 15-16) describes it "the Sketch Engine has come out of the academic research world", and it is today used in linguistics, and languages departments (teaching and research), and in Computational Linguistics and Natural Language Processing. In addition, it is used in language teaching (especially in English language teaching), lexicography, translation and teaching translation, discourse analysis, etc. (Kilgarriff et al. 2014).

Below we will present a widely available non-commercial version of this platform (adapted by JeRTeh) which, in spite of a limited number of tools available, offers a multitude of possibilities for querying and researching (parallel) electronic corpora that consist of learners' production in their mother tongue, and English.

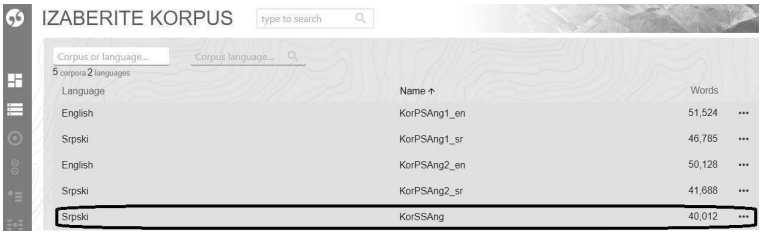


Figure 1. Corpus selection

3.2 Example of *KorSSAng* subcorpus search

First we select one, among the five corpora available, KorSSAng, which consists of argumentative essays written in Serbian, by Serbian students of the English language (Figure 1).

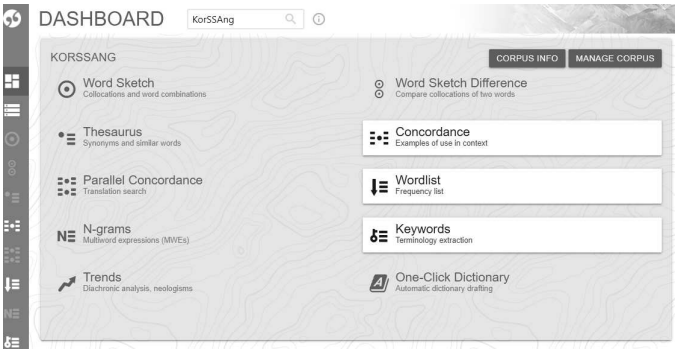


Figure 2. Dashboard

This takes us to the Dashboard (Figure 2) which offers three tools: Concordance, which provides examples of the use (of a term) in context, Wordlist, which generates a frequency list, and Keywords, which can be useful in lexicography (for terminology extraction, etc.).

Corpus Info (Figure 3) can be accessed from the Dashboard as well.

This tab offers the general information about the selected corpus, and provides its description,[7] the tagset used,[8] and the numbers of tokens, words, lemmas, and documents that comprise the corpus. By selecting the *Manage corpus* button (the upper right-hand corner), the corpus can be expanded and modified, by adding texts to the corpus, or by forming new, or altering the existing subcorpora, etc.

Next, having selected the *Concordance* tool, we perform a *Basic* query for the lemma [млад] ('young') (Figure 4), given that the essays have pre-

7. KorSSang

8. 1. N (Noun), 2. A (Adjective), 3. V (Verb), 4. PRO (Pronoun), 5. NUM (Number), 6. PREP (Preposition), 7. CONJ (Conjunction), 8. INT (Interjection), 9. PAR (Particle), 10. ADV (Adverb), 11. PREF (Prefix), 12. ABB (Abbreviation), 13. RN (Roman numeral), 14. PUNCT (Punctuation), 15. SENT (Sentence end marker), 16. ? (Non-Serbian words or suffixes in compounds).

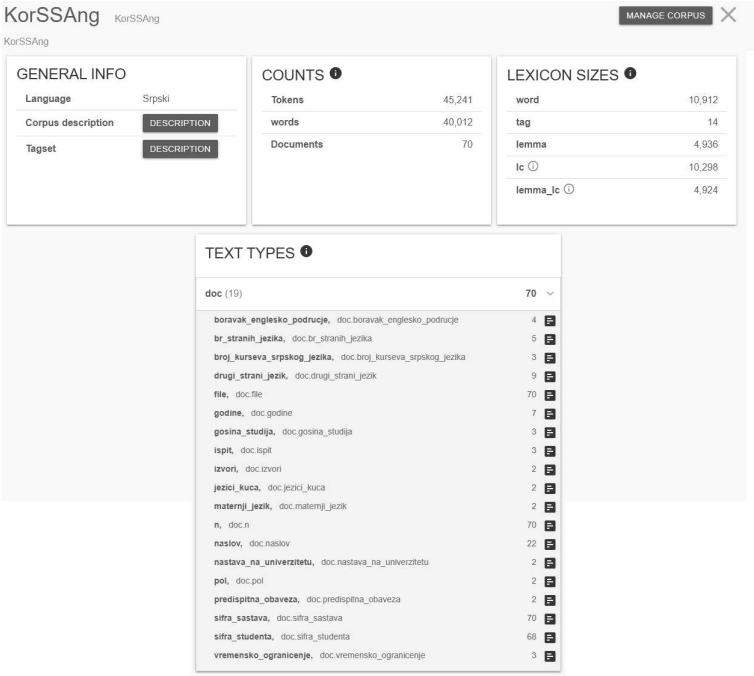


Figure 3. General info about the *KorSSAng* corpus

dominantly been written by the young, hence a significant distribution of this lemma can be expected in this corpus.

The query yields 103 results (such as the uninflected forms of the adjective "млад" ('young'), as well as the use of this adjective for a generic reference denoting young people - "млади" ('the young'), arranged in the order in which they occur in the numbered documents which make up the *KorSSAng* corpus (Figure 5).

The selection of any of the concordance lines offers a wider context of use of the lemma (Figure 6).

The search results can be sorted according to the left, or the right context (i.e. words that appear to the left, or to the right of the lemma), or according to the *Key Word in Context* (KWIC), which is the most common way of sorting concordance lines (Figure 7). The results can be exported in several formats, such as *TXT*, *CSV*, *XLS*, *XML*, whereas the current view can be saved as a *PDF*.

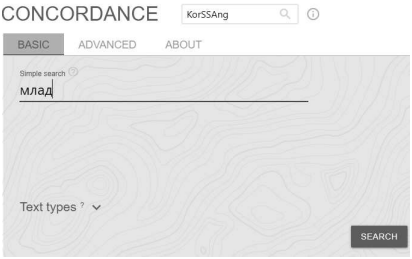


Figure 4. Basic query for the lemma [млад] (‘young’)

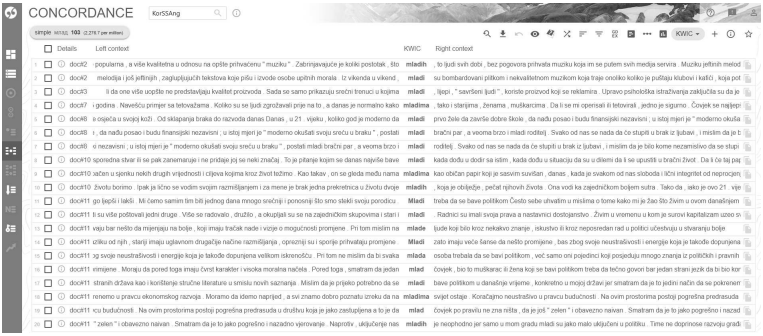


Figure 5. Search results for the basic query of the lemma [млад] (‘young’)

The frequency (Figure 8) and the distribution of the lemma in the entire corpus can be checked as well, along with the words it most frequently collocates with (Figure 9); 1-5 words to the left of the lemma, and 1-5 words to the right (in this case it is the word “људи”[9] (‘people’) with 25 co-occurrences).

Next, we perform an advanced *CQL* query (Context Query Language), which is accessed through the menu in Figure 4, by selecting the tab *Advanced*[10] (Figure 10).

9. By contrast, the query for a noun phrase [млади људи] (‘young people’) in the comparable corpus *LOCNESS* (i.e. its American component, which consists of approximately 150,000 words) yields 12 concordance lines.

10. In addition to CQL queries, advanced searches can be performed for the simple form of the word (which retrieves all the instances of the word that do not include a capital (initial) letter, or for lemmas, phrases, words (regardless of whether they contain capital letters or not), or for special characters (such as punctuation marks, numbers, etc).

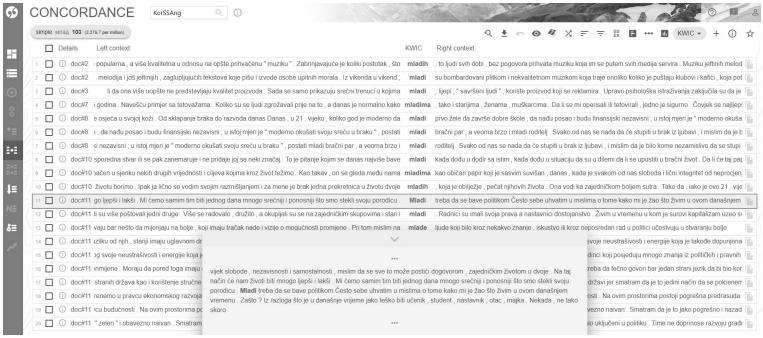


Figure 6. The wider context of use of the lemma [млад] (‘young’)

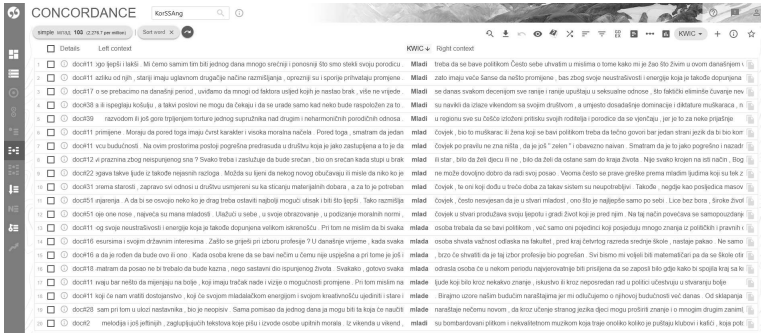


Figure 7. KWIC for the lemma [млад] (‘young’)

For the purpose of this paper we use the CQL Builder, a tool integrated in the Sketch Engine (Figure 11), aimed at generating specific syntax queries for advanced corpus search with the set of tags listed in Footnote 8. In this case we will search for verbs that appear three words to the right of the lemma [млад] (‘young’).

This query yields 94 results (Figure 12). However, before exporting, the concordance lines have to be selected manually in order to single out unique examples, so as to exclude those that are repeated several times due to the specific search range applied {0,3}.

Next we shall present concordance queries in parallel corpora: *KorPSAng1_en* and *KorPSAng1_sr*, and *KorPSAng2_en* and *KorPSAng2_sr* respectively. These corpora can be accessed through the menu in Figure 1 as well.

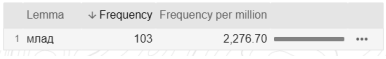


Figure 8. The frequency of the lemma [млад] ('young') in the entire *KorSSAng* corpus

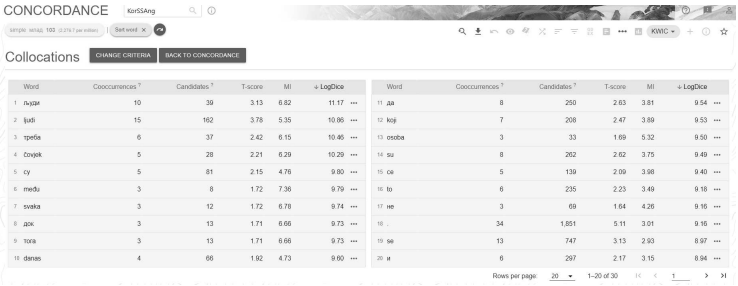


Figure 9. Words which the lemma [млад] ('young') most frequently collocates with

First we select the *KorPSAng1_en* corpus and access the Dashboard (Figure13). This Dashboard differs from the one in Figure 2 2 in that it has an additional tool; Parallel Concordance, since there is a Serbian corpus that *KorPSAng1_en* is paired with (*KorPSAng1_sr*).

Figure 10. Advanced query selection

Then we select the *Parallel Concordance* tool and perform a basic search[11] for the noun phrase "project's collapse" (which is a part of a larger relative clause - *which then led to the project's collapse*) present in one of the source texts in English, to see how it is translated into Serbian 14.

Figure 11. Generating a CQL query `[lemma="млад"] [ ] {0,3}[tag="V.*"]`

The search for the noun phrase "project's collapse" yields 19 parallel concordance lines (each line consisting of an aligned text in English and its translation into Serbian) (Figure 15)

A detailed overview of the results yielded shows that the noun phrase is translated into Serbian in several different ways, predominantly as a noun phrase: *колапс пројекта* (4 examples), *пропаст* (*целог*) *пројекта* (4 examples), *распад пројекта* (3 examples), *пад пројекта*, *гашење пројекта*, *колапс* (*самог*) *филма*, *колапс*, *крај првобитне верзије*, *крах пројекта*, *пропадање пројекта*, and, in one case, it is translated as a part of a clause: *које су довеле до тога* [. . .] *да цео пројекат пропадне.*

We will next perform an advanced CQL query[12] for perfect verb forms in the same parallel corpora; `[lemma="have"] [tag="V.*"]` (Figure 16), since complex verb phrases, more frequently than not, present a stumbling block in the learning of a foreign language.

11. An advanced search for this noun phrase can be performed via the following CQL query as well: `[lemma="project"][word="'"][word="s"][lemma="collapse"].`

12. CQL queries of the English components of the corpus are generated with the Penn Treebank Tagset.

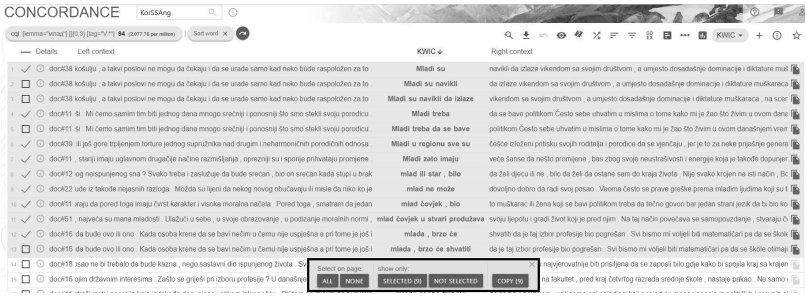


Figure 12. Search results for the CQL query `[lemma="млад"] [ ]{0,3}[tag="V.*"]`

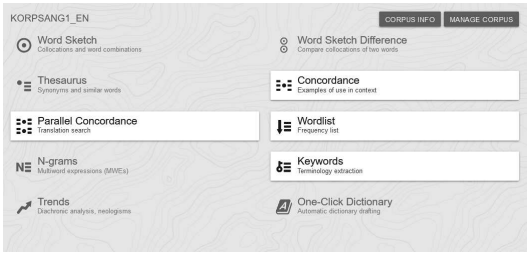


Figure 13. Parallel corpora Dashboard

This search yields 432 results (Figure 17) that include examples of the past, and the present perfect tense, perfect infinitives, and participles.

The search can be further narrowed down using a fine tagset,[13] and, if we want to reduce it to the forms of the past perfect tense, we can do so with the following CQL query: `[word="had"][tag="VVN.*"]`, (where *VVN*

13. VB → verb BE, base form (be), VBD → verb BE, past tense (was, were), VBG → verb BE, gerund/present participle (being), VBN → verb BE, past participle (been), VBP → verb BE, sing. present, non-3rd (am, are), VBZ → verb BE, 3rd person sing. present (is), VH → verb HAVE, base form (have), VHD → verb HAVE, past tense (had), VHG → verb HAVE, gerund/present participle (having), VHN → verb HAVE, past participle (had), VHP → verb HAVE, sing. present, non-3rd (have), VHZ → verb HAVE, 3rd person sing. present (has), VV → verb, base form (take), VVD → verb, past tense (took), VVG → verb, gerund/present participle (taking), VVN → verb, past participle (taken), VVP → verb, sing. present, non-3rd (take), VVZ → 3rd person sing. present (takes).

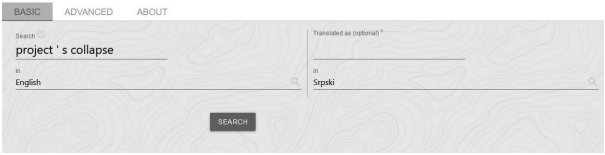


Figure 14. Basic search for the noun phrase "project's collapse"

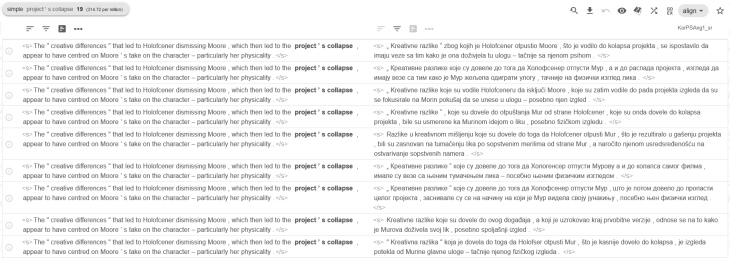


Figure 15. Basic search results for the noun phrase "project's collapse"

is the tag for the past participle of the main verb). This query yields 168 results (Figure 18).

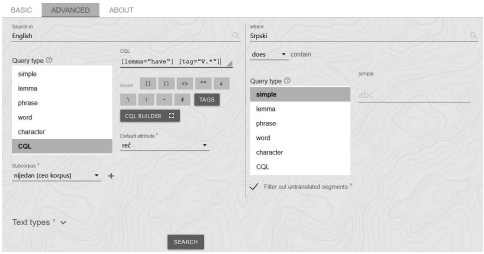


Figure 16. The advanced CQL query for perfect verb forms; `[lemma="have"] [tag="V.*"]`

Finally, through the menu in Figure 1, we select the *KorPSAng2_sr* corpus and perform an advanced CQL query for the noun phrase "istrošene svađalačke snage" `[lemma="истрошен"] [lemma="свађалачки"] [lemma="снага"]` to search for parallel concordances (Figure 19).

This query yields 32 results. What follows are several examples of various complexity: dissipated energy for quarrelling, a drained argumentative spirit, spent fighting endurance, the wasted energy on fights, used up strength for

quarreling, used up strength for falling out, worn out quarrelsome energy, wasted fighting energy, wasted argumentative energy, a drained will to fight, a small fighting force, etc.

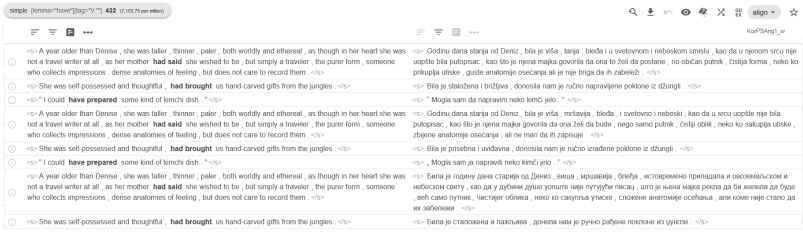


Figure 17. CQL query `[lemma="have"] [tag="V.*"]` search results

4 Conclusion

In this paper, we have briefly described the steps of the creation of the Corpus of English-studies Students (*KorSAng*), and presented the possibilities of its search. It is known that the lack of student corpora in particular is a handicap for researchers in the field of applied linguistics, and we hope that our efforts to overcome this problem will result in greater popularity of corpus linguistics as a research method in this area, especially in English studies. The results of the use of the *KorSAng* corpus so far have been presented in the following works: (Šućur 2020; Spajić and Suknović 2019; Tomović and Stefanović 2019; Марковић and Станковић, у штампи), and a doctoral dissertation, part of the corpus of which is based on *KorSAng*. The scientific potential of this corpus can be expected in many scientific and professional papers, monographs and scientific research projects. We will add that the plan is to create a second version of the *KorSAng* corpus, with an expanded number of essays and translations, and with the possibility of integrating new software tools that will provide easier search and more comfortable use of the corpus to end users.

Acknowledgment

This paper is based on research conducted within two national projects: *Scientific potentials of annotated student corpora in applied linguistics*

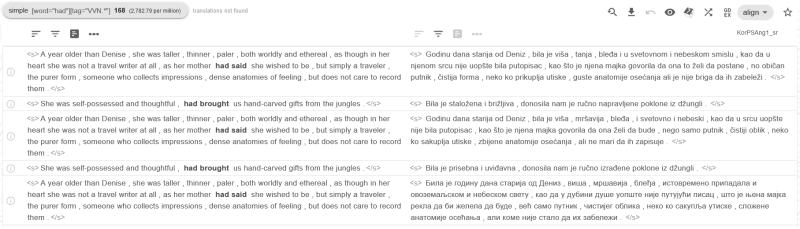


Figure 18. Search results for the past perfect tense via a CQL query `[word="have"] [tag="VVN.*"]`

[Naučni potencijali anotiranih učeničkih korpusa u primijenjenoj lingvistici], 19.032/961-135/19 and *Phraseological competence of Serbian speakers of English through the prism of contrastive analysis of interlanguage* [Frazeološka kompetencija srpskih govornika engleskog kroz prizmu kontrastivne analize međujezika], 19/6-020/961-46/18. The projects were financially supported by the Ministry for Scientific and Technological Development, Higher Education and Information Society, Banja Luka. We thank *The Society for Language Resources and Technologies – JeRTeh* for the cooperation in the project (19.032/961-135/19), especially Professor Ranka Stanković. Special thanks go to project coordinator, Professor Jelena Marković, for providing us with the opportunity to cooperate in the projects mentioned, as well as for her continuous support in the form of encouragement, advice, motivation and expertise.

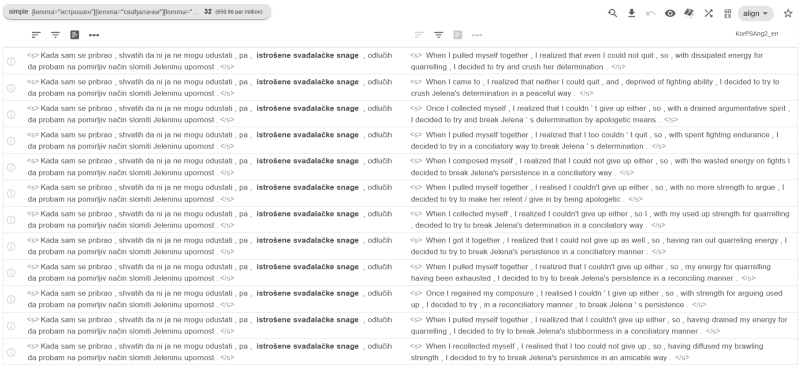


Figure 19. Basic search results for the noun phrase “istrošene svađalačke snage”

References

CEO. 2015. *Општи стандарди постигнућа за крај општег средњег и средњег стручног образовања и васпитања у делу општеобразовних предмета.* Београд: Завод за вредновање квалитета образовања и васпитања.

Granger, Sylviane. 2002. "A bird's-eye view of learner corpus research." *Computer learner corpora, second language acquisition and foreign language teaching* 6:3–33.

ISO24616. 2012. *Language resources management — Multilingual information framework.* International Standard Organization.

Kilgarriff, Adam, Vıt Baisa, Jan Bušta, Miloš Jakubıček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychl, and Vıt Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

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

Kunilovskaya, Maria, and Marina Koviazina. 2017. "Sketch engine: A toolbox for linguistic discovery." *Jazykovedny Casopis* 68 (3): 503.

Radić-Bojanić, Biljana B. 2019. "NEODREĐENA ZAMENICA ONE U PISANJU KOD NEIZVORNIH GOVORNIKA ENGLESKOG JEZIKA." *Годишњак Филозофског факултета у Новом Саду* 44 (2): 39–52. https://doi.org/10.19090/gff.2019.2.39-52.

Radonja, Minja S. 2019. "The use of interactive metadiscourse in Serbian students." *Радови Филозофског факултета: Часопис за хуманистичке и друштвене науке* 8 (21). https://doi.org/10.7251/FIN1921121R.

Spajić, Sonja, and Mina Suknović. 2019. "The Choice of Lexemes According to Their Frequency in Translation into L2." *Komunikacija i kultura online* 10 (10): 104–119. https://doi.org/10.18485/kkonline.2019.10.10.6.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of The 12th Language Resources and Evaluation Conference,* 3954–3962.

Šućur, Srđan. 2019. "Distribucija frazalnih glagola u pisanju na engleskom kao stranom kod srbofonih govornika." *Komunikacija i kultura online* 10 (10): 120–143. https://doi.org/10.18485/kkonline.2019.10.10.7.

Šućur, Srđan. 2020. "REVERSE TRANSFER IN ADULT SERBIAN EFL LEARNERS'WRITING1 2-A CORPUS BASED STUDY." *BEYOND HERMENEUTICS,* 141.

TMX. 2005. "Translation Memory eXchange (TMX) 1.4b Specification." Accessed 1.08.2021. https://www.gala-global.org/knowledge-center/industry-development/standards/lisa-oscar-standards.

Tomović, Nenad, and Sofija Stefanović. 2019. "Uticaj L2 i leksički i leksičko-sintaksički kalkovi u prevodu. Studija slučaja." *Komunikacija i kultura online* 10 (10): 144–154. https://doi.org/kkonline.2019.10.10.8.

Töny, Luzius. 2016. "Corpora als Ressourcen für die maschinelle Übersetzung." Accessed 17.04.2016. https://swanrad.ch/downloads/mt_1.pdf.

Utvić, Miloš, Ranka Stanković, and Ivan Obradović. 2008. "Integrisano okruženje za pripremu paralelizovanog korpusa." *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen,* 563–578.

Vitas, Duško. 2010. "Resursi i metode za obradu srpskog – stanje i perspetive." In *Srpska lingvistika/Serbissche Linguistik, Eine Bestandsaufnahme, Studies of Language and Culture in Central and Eastern Europe (SLCCEE,* edited by Biljana Golubović and Cristian Voß, 7:257–277. München: Verlag Otto Sagner.

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

Андоновски, Јелена. 2021. "Паралелни корпуси у Србији — могућности за паралелно проналажење информација на два или више језика" [in serbian]. 3, *Библиотекар* 63 (1): 51–74. ISSN: 0006-1816. https://doi.org/10.18485/bibliotekar.2021.63.1.3.

Марковић, Јелена. 2017. "Лични метадискурс у писању код изворних и неизворних говорника енглеског језика." *Филолог–часопис за језик, књижевност и културу,* no. 15, 44–60. https://doi.org/10.21618/fil1715044m.

Марковић, Јелена. 2018. "Употребе глагола make у писању на енглеском језику као страном код изворних говорника српског језика (корпуснолингвистичка анализа)." *Зборник матице српске за филологију и лигвистику* 61 (1): 165–180. https://www.maticasrpska.org.rs/stariSajt/casopisi/ZMSFL_61_1.pdf.

Марковић, Јелена. 2019. *Кроз призму контрастивне анализе међујезика.* Филозофски факултет.

Марковић, Јелена. 2020. "Концесивни конектори though и however у писању на енглеском језику код изворних и неизворних говорника." *Филолог–часопис за језик, књижевност и културу,* no. 21, 13–35. https://doi.org/10.21618/fil2021013m.

Марковић, Јелена, and Ранка Станковић. у штампи. "Ја/ти/ми/ви у дискурсној компетенцији у светлу контрастивне анализе међујезика." *Методички видици.*

Ристовић, Зоран. 2012. "Од корпуса до учионице: примена паралелизованих текстова у настави енглеског језика у основној школи." *ИНФОтека* 13 (2): 52–66.

Ристовић, Зоран. 2016. "Кумулативни ефекти експлоатације вишејезичних корпуса у настави страних језик." PhD diss., Универзитет у Београду, Филолошки Факултет.

Шућур, Срђан Р. 2020. "Корпус као оруђе за проницање тајни међујезика." *Радови Филозофског факултета: Часопис за хуманистичке и друштвене науке,* no. 22, https://doi.org/10.7251/RFFP2022321S.

Wiki-Librarian: A Project to Train Librarians and Students to Work with Wikipedia

UDC 023-051: [030:004.738.5
DOI 10.18485/infotheca.2021.21.1.3

ABSTRACT: Wiki-librarian is a multi-year project to train librarians and students of librarianship and information science to use wiki tools, including writing articles on Wikipedia. The project has existed since the beginning of 2015, and the University Library "Svetozar Marković" and is logistically and financially supported by Wikimedia Serbia. The part of the project that deals with the training of librarians is officially accredited. Moreover, the project has expanded its scope to work with students, as well as with librarians outside the training process, through editathons, digitization of content, participation in other wiki activities such as 1Lib1Ref, Wikipedian in Residence. Project activities were presented at several conferences and partially in several publications. On the one hand, better awareness of librarians of wiki software IT capabilities as well as methods for improved presentation of their knowledge in the digital wiki environment are achieved by this project, while on the other hand these activities significantly increase the textual resources of Serbian Wikipedia and their quality.
KEYWORDS: Wikipedia, training of librarians, Wikipedia in education, references on Wikipedia.

PAPER SUBMITTED: 30 May 2021
PAPER ACCEPTED: 19 July 2021

Đorđe S. Stakić
djordje.stakic@ekof.bg.ac.rs
University of Belgrade
Faculty of Economics
Belgrade, Serbia

Aleksandra K. Popović
popovic@unilib.rs
Oja Krinulović
okrinulovic@unilib.rs
University Library
"Svetozar Marković"
Belgrade, Serbia

1 Introduction

The online encyclopedia Wikipedia[1] has established itself as an indispensable information resource that seeks to gather and make available widespread knowledge contributed by many people especially in the era that produced Internet as the main medium that connects different parts of the world in real time. Keeping pace with this trend, librarians are modernizing their work and the access to information sources and users. Wiki-librarian project organized by the University Library "Svetozar Marković" from Belgrade and Wikimedia Serbia[2] was created as an expression of readiness to find the best way that unites Wikipedia and librarians. After several years this project has produced a large number of trained librarians and students of librarianship to work on Wikipedia, which resulted in numerous written and updated articles, posted references, created and digitized photographs, scanned old books.

Since the very beginning of Wikipedia there were questions regarding the attitude of librarians towards Wikipedia. According to the research (Luyt et al. 2010), this attitude was mostly positive. According to (Snyder 2013), librarians are aware of the negative rather than the positive aspects of Wikipedia; therefore they use it more for personal purposes than for research or to help the users. The research (Vetter 2014) describes how music composition students at the University of Ohio used special library collections to create articles for Wikipedia, which increased students' motivation and public awareness of the library resources. The review paper (Okoli et al. 2014) analyzes 99 scientific papers that deal with Wikipedia to some extent and examine the attitude of these works towards Wikipedia readers. Common Wikipedia users include researchers, librarians and students. In (Jemielniak and Aibar 2016) the perception and distrust of the academic community towards Wikipedia are analized while the possible scenarios of peaceful coexistence are stated. The (Faletar Tanacković, Đurđević, and Badurina 2015) paper presents research conducted at the Faculty of Philosophy in Osijek on the experiences and attitudes of students and teachers on the use of Wikipedia in academic environment. It showed that students tend to have better opinion about Wikipedia than teachers do, and that 19.8% of teachers explicitly forbid students from using Wikipedia.

The paper (Kousha and Thelwall 2017) is an interesting study on the citation of Wikipedia articles in scientific papers and monographs. The sam-

1. English Wikipedia, Serbian Wikipedia, accessed July 1st 2021.
2. Wikimedia Serbia, accessed July 1st 2021.

ple included 302,328 scientific papers and 18,735 monographs indexed in Scopus[3] for the period from 2005 to 2012. Only 5% of scientific papers had a Wikipedia article among their references, while in the monographs their percentage was 33%. In the review papers, that percentage is 8%. It means that Wikipedia is more used as a as a source in writing books than it is the case with scientific papers.

The (Lewoniewski, Węcel, and Abramowicz 2017) research on 7 major Wikipedias (by number of articles) and over 10 million of their articles showed that a significant number of references are repeatedly referenced on many different Wikipedias. Thus, the English and German Wikipedias had the highest number of common references - 345202, which is understandable because they are the biggest encyclopedias from that collection. If articles presented on at least 5 Wikipedias are taken into consideration, then the highest number of common references was between the English and Russian Wikipedias. According to (Singh, West, and Colavizza 2021) research on a sample of 29.3 million citations from 6.1 million Wikipedia articles in English, it was found that 6.7% of Wikipedia articles cite at least one paper from a journal with an associated DOI,[4] while Wikipedia cites about 2% of the total articles that are indexed in the Web of Science[5] with associated DOI. The (Pooladian and Borrego 2017) study shows lack of standardization while citing references on Wikipedia with data often missing or having errors. That research showed that only 3% of scientific articles from information and library sciences published in the period 2001-2010 were cited on Wikipedia until the end of 2016. The sample included 26542 scientific papers from the category "Information Science and Library Sciences" indexed in the Web of Science.

2 The context of creation of the project

Wikimedia Serbia was founded in December 3rd 2005 as a local chapter of the Wikimedia Foundation.[6] Its activities are largely focused on cooperation

3. Scopus is a scientific citation database created in 2004; Scopus, accessed July 1st, 2021.

4. DOI is an abbreviation for "Digital object identifier", used to uniquely identify scientific papers and other digital resources. DOI, accessed July 1st, 2021.

5. Web of Science is a leading scientific citation database; Web of Science, accessed July 1st, 2021.

6. Information on Wikimedia affiliates, thematic organizations, and user groups is available at Wikimedia chapters, accessed July 1st, 2021.

with educational institutions, primarily with colleges but also with the secondary schools. In December 2005, the idea about students getting involved in writing articles on Wikipedia was presented within the seminar at the Department of Computer Science and Informatics at the Faculty of Mathematics, University of Belgrade. The founder and main promoter of this idea was one of the authors of this text, Đorđe Stakić. The idea was realized in the upcoming year after two faculties of the University of Belgrade, The Faculty of Physical Chemistry and The Faculty of Philology took part in the educational activities of Wikimedia Serbia by writing articles on Wikipedia instead of classic seminar papers. The first article posted on Serbian Wikipedia was ENIAC,[7] dedicated to the first digital electronic computer. It was published by a student of the Faculty of Philology on April 18th 2007 within the course "Internet and Web Technology" held by Professor Cvetana Krstev.

Other faculties and high schools were involved in the further development of the educational program of Wikimedia Serbia. In April 2012, the Academic Board of Wikimedia Serbia was formed as a group of members that focused on the development and implementation of educational activities related to Wikipedia. More details about wiki technology, its development and importance, as well as the pioneering endeavors of the development of the educational program on Wikipedia are available in the paper (Stakić 2009). Research results of students' attitudes towards the use of Wikipedia as a teaching tool performed at the Faculty of Pedagogy in Vranje of theUniversity of Niš are presented in the paper (Stakić et al. 2021). This research was made on a sample of 203 students and it showed that they positively evaluate the use of Wikipedia in teaching and that they prefer writing articles on Wikipedia rather than writing classic seminar papers.

The University Library "Svetozar Marković" in Belgrade was officially opened on May 24th 1926. Today, it is one of the biggest scientific libraries in Southeast Europe with about 1.7 million publications. The training programs for librarians have been implemented within its activities from 2004. These training programs got official accreditation in 2014, see (Popović, Stolić, and Stanišić 2019). The University Library shows a great readiness and will to cooperate with other organizations that are committed to common goals. As part of that cooperation, a panel discussion "Creative Commons - Creativity and the Knowledge Society" was held at the University Library on December 12th 2013, in a joint organization of Wikimedia Serbia and the

7. ENIAC in Wikipedia, accessed July 1st, 2021.

University Library.[8] Đorđe Stakić held a presentation on "Wikimedia Serbia and the University of Belgrade in the development of open knowledge" and presented the educational program and work of the Wikimedia Serbia Academic Board. Aleksandra Popović and Dragana Stolić held a presentation "Creative Commons, licenses and e-thesis of the University of Belgrade". The panel discussion initiated the idea for the cooperation of the University Library and Wikimedia Serbia in the field of educational programs as well as to design a training program for librarians to work on Wikipedia.

3 Development of the Wiki-librarian project

Following the initial plans and agreements between the University Library and Wikimedia Serbia, the realization of the internal workshop for training of librarians of the University Library soon started. The librarians were the main bearers of future project having in mind that they were responsible for its realization. This training took place on January 30th 2014, in the premises of the library. During the summer of 2014, a Project Proposal for the Wiki-librarian was sent to Wikimedia Serbia which approved funds for its implementation as part of its annual plan for 2015. At about the same time, the project also applied at the National Library for an accredited program of continuous professional development of librarians. The project application was submitted by Aleksandra Popović, Oja Krinulović and Đordje Stakić, as authors of the project, as well as the University Library "Svetozar Marković" and the Faculty of Mathematics of the University of Belgrade as institutions in charge of the project. At the end of the year, the project was officially accredited for 2015. Later on, the project was accredited three more times, for a year (2016), for the two upcoming years (2017-2018) and finally for the three upcoming years (2019-2021). Wikimedia Serbia was also approved for the project for each subsequent year, starting from 2015 to 2021, with the corresponding annual budget which would cover the project costs.

In order to prepare for the project, Wikimedia Serbia organized the first one-month program, Wikipedian in Residence, realized within the University Library in December 2014. Jovana Milošević was a student hired to hold the

8. The agenda of a panel discussion "Creative Commons - Creativity and the Knowledge Society", access date May 30th 2021.

workshops for librarians, write new articles, digitize and post photos under a free license, to the Wikimedia Commons[9] for this project

The realization of accredited workshops for librarians began in 2015, and it was conducted by accredited lecturers, members and project implementers. The workshops were realized as two-day events with duration of three hours each, and according to the accreditation, they carry a total of 6 points. On the first day, training for work on Wikipedia is realized in the form of lectures and workshops. The second day is about writing and posting articles on Serbian Wikipedia. Librarians are encouraged to prepare material for the article, which is than uploaded and adapted to wiki syntax.[10] An important part of the whole process of writing articles is citing references from relevant sources, which contributes to the verifiability of articles on Wikipedia, that is very common to the librarians. Upon completion of the training, all participants receive certificates, and the results in the form of written articles remain on Wikipedia. In this way, all training participants have a certain benefit and motivation to participate in the project. Having in mind that the topics are open, librarians have the opportunity to write about their libraries, sights from where they come from and else that is relevant to Wikipedia and is over threshold of notability.[11] The enriched content of Serbian Wikipedia has thus contributed to higher visibility of libraries in the search engine results, as well as greater affirmation of the librarian profession in general. In this way, the librarians themselves have the opportunity to get to know and connect with each other as well as to get a chance to work on something of greater significance.

Apart from the accredited two-days workshops, for the interested librarians, additional workshops were organized for further training in certain activities, but also to ensure a larger number of written articles. That led to another 7 workshops during 2015. The following activities within the framework of this project were edit-a-thons. They started as live, few hours one-day activities, which gathered the interested participants – librarians who write and post articles on Wikipedia. Experienced librarians were helping the beginners, and in general, the participants were helping each other to manage with the new and less familiar things. Since the activities were organized in

9. Wikimedia Commons is a repository of media files under free licenses; Wikimedia Commons, accessed July 1st 2021.

10. Wiki syntax is the procedure for text marking in the MediaWiki software used on Wikipedia; formatting rules, accessed July 1st 2021.

11. Wikipedia has guidelines on notability; Wikipedia:Notability, accessed July 1st 2021.

the Library, the necessary literature was at hand which made the creation of the articles much easier. Three edit-a-thons were organized during 2015: "Open science", "150 years from the birth of Jovan Cvijić" and "Librarians together on Wikipedia". During 2016, two edit-a-thons were organized on the topic "Open access scientific journals" with a total of 4 workshops. Librarians from the journal publishing institutions and representatives from the open-access scientific journals published in Serbia were invited. Since the response was unexpectedly big, the edit-a-thon was repeated with several workshops realized. In that way, Wikipedia became richer for about 60 articles in this field, which significantly contributed to the accessibility of the information about these magazines to a wider circle of readers. During 2017, the edit-a-thon "Open access scientific journals" was repeated through two workshops. The activities continued in 2018, when there were two edit-a-thons about scientific journals and another one called "My Street". Since a large number of librarians from various cities went through the training, this marathon had a response within many of the previous participants in accredited workshops. For that reason, it was repeated twice in 2019.

Starting in 2016, within the project Wiki-librarian, the University Library also holds an internship for students of Librarianship and information science from the Faculty of Philology, University of Belgrade. This internship is related to the course Digital Text II. Within the internship, students go through trainings, practical workshops, they write articles and in the end, they get points which form the final grade for the Digital Text II course. This is very useful for students to experience the practice of a large and important library such as the University library and get acquainted with the work in the library, with the process "from the inside", as well as with the employed librarians. These are mostly multi-day trainings and result in new articles on Wikipedia.

Starting from 2018, participation in the global 1Lib1Ref[12] campaign was added to the list of above mentioned activities. This campaign exists with purpose to motivate the librarians to contribute professionally to Wikipedia by adding references to the articles. The campaign itself started on January 15th, on Wikipedia's birthday, and lasted for about 20 days. Apart from the librarians, other users are also welcome to participate in referencing in this campaign.

The pandemic in March 2020 affected the activities related to this project. It was no longer possible to organize live workshops for either librarians or

12. The 1Lib1Ref (abbreviated for "one librarian, one reference") global campaign; 1Lib1Ref, accessed July 1st 2021.

students, just as it was no longer possible to have live edit-a-thons. However, that did not stop the activities of the project, which by its nature, managed to adapt and to realize a significant part of the activities online, and in that way achieve the planned metrics. Thus, in 2020, the student internship was realized through sending presentations, brochures and video materials with e-mail consultations, while in 2021 the student training was held online via Zoom,[13] with subsequent consultations by e-mail. During 2020, three multi-day online edit-a-thons were held: "Serbian humorous-satirical periodicals of the 19th and 20th century", "My street", "Press of my region", as well as the 2021 edit-a-thon "Literature for children". The diversity of these topics contributed to the good response from librarians. During 2020, attention was paid to the digitization of publications within the project, and during 2020 and 2021, a significant part of the activities was related to the global campaign 1Lib1Ref.

4 Achieved results

As part of the Wiki-librarian project, three one-month internship programs within the frame of Wikipedian in Residence were held in December 2014, July 2016 and March 2019. The first internship resulted in 18 new articles on Wikipedia, 102 uploaded photos and 5 scanned books with a total of 4554 pages. During this internship, 8 internal workshops were held in the University Library for 10 employees.[14] During the second internship, 8 books (2585 pages) were digitized and placed on Wikimedia Commons.[15] The third internship contributed with 16 new, 13 supplemented, 13 edited articles, 158 photographs and 10 books (2607 pages).[16]

A great number of librarians from different cities in Serbia went through accredited Wiki-librarian workshops. During 2015 and 2016, there were 6 such two-day workshops, which were attended by about 70 librarians per year with approximately as many written or updated articles. During 2017 there were 7 workshops, in 2018 there were three workshops and during 2019 there were four realizations of the accredited seminar. Most workshops were realized in Belgrade, with over 100 librarians attendees. In addition,

13. Zoom is video conferencing software that expanded during the 2020 pandemic; Zoom official website, accessed July 1st, 2021.

14. Details of this internship are available at the website: Details of the first internship, access date July 1st 2021.

15. Details of the second internship, access date July 1st 2021.

16. Details of the third internship, accessed July 1st 2021.

accredited workshops were held in 13 other cities: Užice, Novi Sad, Banja Vrujci, Čačak, Smederevo, Vranje, Valjevo, Leskovac, Obrenovac, Topola, Kragujevac, Jagodina and Kruševac. Holding the workshops in various cities throughout Serbia gave significant contribution to the decentralization of the activities. There were 8 to 19 librarians from 3-17 libraries per each of these workshops outside Belgrade. In these metrics, each library employee that passed the training was titled librarian. According to (Popović, Stolić, and Stanišić 2019), during 2015-2018, a total of 22 accredited Wiki-librarian workshops were held, with a total of 298 participants. During 2019, another 42 participants passed the training through 4 accredited workshops. Taking into account not only the places the seminar was held at, but where the participants were from in general, we come to the number of 39 cities and municipalities in Serbia. The participants of the seminar were from different types of institutions, as shown in Table 1. Apart from these official ones, there were additional workshops for librarians, 7 during 2015, 3 during 2016, and 4 during 2017, 2018 and 2019. Apart from the trainings, the contribution from workshops were updated articles on Wikipedia. Since 2020, as mentioned before, there has been no live workshops due to coronavirus pandemic.

Institutions	**Number**
Public libraries	34
Archives	2
Museums	6
Schools	15
Societies - associations	11
Higher education institutions	77

Table 1. Data on the number of institutions that attended the Wiki-librarian course from 2015 to 2021

If it was necessary for the participants, workshops and lectures were also held on Wikipedia during the edit-a-thons. Whether the edit-a-thons were held live or online, they resulted in increased number of new and updated articles. The following edit-a-thons were among the most successful:

– "Librarians together on Wikipedia" in 2015, with 44 new and 6 updated articles,

– "Open access scientific journals" in 2016 with a total of 58 new and 1 updated article, while in 2017 there were 25 new and 3 updated articles,
– "My street" in 2018, with 37 new and 1 updated article, while during 2019, 44 new articles were written.

Particularly successful were the multi-day online edit-a-thons in 2020 and 2021, which metrics are:

– "Serbian humorous-satirical periodicals from the 19th and 20th century": 71 new, 5 updated
– "My street": 59 new, 5 updated
– "Press of my region": 91 new, 6 updated
– "Literature for children": 76 new, 12 updated.

Some of these edit-a-thons made significant contribution considering the free files uploaded to the Wikimedia Commons. Three edit-a-thons from 2020 were especially significant, with a total of about 800 uploaded files.[17]

Through the internships during 2016-2021, the trainings of Wiki-librarians were attended by a total of 143 students and as a result an average of between 20 and 30 articles were received per year, including a significant number of uploaded free files, mostly photographs. For example, as part of a student internship in 2021, 36 files were uploaded to Wikimedia Commons.

Since Wiki-librarian inclusion in the 1Lib1Ref campaign in 2018, the achieved results concerning the upload of references were getting better. Thus, in 2018, 786 references were added, in 2019 another 1672, in 2020 another 4052, while in 2021, 17112 references were added. On a global level, the community of Serbian Wikipedia, according to the number of references added within 1Lib1Ref campaign, took the third place in 2019, the second place in 2020, and the first place in 2021. In 2021 (which was top rated according to the number of added references), among the 10 users who added most references on the global level, 5 of them were from the Serbian Wikipedia.[18]

Since 2015, through all the activities held in the Wiki-librarian project in total, several hundred articles have been created or updated on Serbian Wikipedia every year.

The Wiki Librarian project was presented at several international and domestic conferences, as well as at several local gatherings of librarians and

17. Activities and results of Wiki-librarians for 2020, access date July 1st 2021.
18. Local results of the 1Lib1Ref campaign for 2021, access date July 1st 2021.

Wikipedians organized on various occasions. In this way, greater dissemination of the project was achieved, several posters and articles were published, which all led to some new collaborations. Among the published articles, (Popović, Ševkušić, and Stakić 2015; Krinulović, Stijepović, and Stakić 2015; Antonić, Živanović, and Popović 2017; Popović et al. 2018) stand out. Posters from the international conferences INFORUM - 22nd Annual Conference on Professional Information Resources (Prague 2016), INFORUM - 23rd Annual Conference on Professional Information Resources (Prague 2017), 10th International Conference of the European Guidelines for Cooperation of Libraries, Archives and Museums (Sarajevo 2018) are attached as photos (figures 1, 2, 3).[19] In the paper (Obradović et al. 2020), the Wiki-librarian is mentioned as one of the Wikimedia Serbia projects, while in the section on the application of Creative Commons license in Serbia, the University Library "Svetozar Marković" was mentioned along with the number of faculties.

The importance of Wiki-librarian project is reflected in the fact that it enabled other libraries to participate in the activities related to Wikipedia through the release and digitization of content, as well as through the realization of the Wikipedian in Residence project. After the University Library, the most successful cooperation was achieved with the Belgrade City Library, starting from 2019 where, in addition to library-type projects, Milica Buha developed the Wiki senior project intended for training and motivating older participants, retirees, to work on Wikipedia.

Another important aspect of Wiki-librarians is help in overcoming the gender gap on Wikipedia (Ratković and Madžarević 2021). According to various surveys which were conducted on a sample of Wikipedia editors, the share of women among them is up to one-sixth or one-fifth at best, see (Graells-Garrido, Lalmas, and Menczer 2015; Maljković 2015). The Wiki Librarian project contributes to a significant increase in the number of women on Wikipedia since the share of women librarians and librarianship students is reversed in relation to the Wikipedia average, with a majority of 80 percent and higher. While this is not an essential item for the project itself, this fact has its value and weight for the bigger picture. The members of the Wiki-librarian project themselves have been participating in the global WikiGap[20] campaign since 2018, which is usually realized in the form of an edit-a-thon on Wikipedia and aims to reduce the gender gap on Wikipedia.

19. The first two posters are also available at the links Poster inforum 2016 and poster inforum 2017, accessed July 1st 2021.

20. WikiGap, accessed July 1st 2021.

At the end of 2017, Wikimedia Serbia declared the Wiki-librarian project as its best project for the year. In addition, Wikimedia Serbia awarded Aleksandra Popović the annual "Branislav Jovanović" award for 2020, as stated in the explanation of the award "For successfully running the Wiki-librarian project and the overall work and Serbian Wikipedia page revision".

5 Conclusion

Digitization of the library fund provides access to different types of content. Therefore the Wiki Librarian project was created and persisted for many years. The leading idea of the project is to enrich Wikipedia, as the most important project of the Wikimedia Foundation, with the highest possible quality content. Libraries hold invaluable treasure that is not easily accessible to everyone. Uploading content to Wikimedia Commons is an ideal way to promote works with expired copyrights or those whose authors want to provide open access. Librarians – as a well-connected community, whose members have access to reference literature in their workplaces, are skilled in finding valid and verifiable information, familiar with the rich funds and collections of their institutions and have experience in training users of different ages and levels of education so they are the perfect choice for Wikipedia editors. By relying on their expertise, librarians can make significant contributions to better data structuring and more efficient indexing of content through Wikipedia projects.

In seven years of existence, the Wiki Librarian project has trained several hundred librarians from all over Serbia to work on Wikipedia through different activities. In addition, several generations of librarianship students at the Faculty of Philology participated in the project. Numerous edit-a-thons were realized within the project, which significantly enriched the content of Serbian Wikipedia. In that way, several thousand articles were posted or updated. In addition, a large number of publications have been digitized and placed on the Wikimedia Commons, where, apart from them, there is also a large number of photographs created and posted by project participants. Particularly successful activity of Wiki-librarians was the participation in the global campaign 1Lib1Ref, in which, during 2021, the community of Serbian Wikipedia entered the largest number of references on a global level. The project contributes to reducing the gender gap on Wikipedia, because due to the circumstances, the vast majority of librarians who passed the project are women. So far, activities of the project are partially presented in several

publications and poster presentations from conferences. These activities are presented in this paper in more comprehensive and studious manner.

References

Antonić, Sanja, Igor Živanović, and Aleksandra Popović. 2017. "Altmetrics - Useful Way for Scientific Communication." Paper presented at the INFORUM 2017: 23rd Annual Conference on Professional Information Resources, May 30–31, 2017, Prague. https://www.inforum.cz/proceedings/2017/24/.

Faletar Tanacković, Sanjica, Anja Đurđević, and Boris Badurina. 2015. "Wikipedija u akademskom okruženju: stavovi i iskustva studenata i nastavnika." *Libellarium: časopis za povijest pisane riječi, knjige i baštinskih ustanova* 8 (2): 161–199. https://doi.org/10.15291/libellarium.v8i2.234.

Graells-Garrido, Eduardo, Mounia Lalmas, and Filippo Menczer. 2015. "First women, second sex: Gender bias in Wikipedia." In *Proceedings of the 26th ACM conference on hypertext & social media,* 165–174. https://doi.org/10.1145/2700171.2791036.

Jemielniak, Dariusz, and Eduard Aibar. 2016. "Bridging the gap between wikipedia and academia." *Journal of the Association for Information Science and Technology* 67 (7): 1773–1776. https://doi.org/10.1002/asi.23691.

Kousha, Kayvan, and Mike Thelwall. 2017. "Are Wikipedia citations important evidence of the impact of scholarly articles and books?" *Journal of the Association for Information Science and Technology* 68 (3): 762–779. https://doi.org/10.1002/asi.23694.

Krinulović, Oja, Mile Stijepović, and Đorđe Stakić. 2015. "Bibliotečki doprinos "srpskoj Vikipediji"." *Glas biblioteke* 21:21–28. https://cacak-dis.rs/izdanja/casopisi/glas-biblioteke/.

Lewoniewski, Włodzimierz, Krzysztof Węcel, and Witold Abramowicz. 2017. "Analysis of references across Wikipedia languages." In *International Conference on Information and Software Technologies,* 561–573. https://doi.org/10.1007/978-3-319-67642-5_47.

Luyt, Brendan, Yasmin Ally, Nur Hakim Low, and Norah Binte Ismail. 2010. "Librarian perception of Wikipedia: Threats or opportunities for librarianship?" *Libri* 60:57–64. https://doi.org/10.1515/libr.2010.005.

Maljković, Filip. 2015. *Izveštaj o rodnoj strukturi na Vikipediji na srpskom jeziku.* Wikimedia Commons, April 27th 2015. https://commons.wikimedia.org/wiki/File:Rodna_struktura_na_Vikipediji_na_srpskom_jeziku.pdf.

Obradović, Ivan, Ranka Stanković, Marija Blagojević, and Danijela Milošević. 2020. "Open Educational Resources in Serbia." In *Current State of Open Educational Resources in the "Belt and Road" Countries,* 175–194. https://doi.org/10.1007/978-981-15-3040-1_10.

Okoli, Chitu, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2014. "Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership." *Journal of the Association for Information Science and Technology* 65 (12): 2381–2403. https://doi.org/10.1002/asi.23162.

Pooladian, Aida, and Ángel Borrego. 2017. "Methodological issues in measuring citations in Wikipedia: a case study in Library and Information Science." *Scientometrics* 113 (1): 455–464. https://doi.org/10.1007/s11192-017-2474-z.

Popović, Aleksandra, Oja Krinulović, Mile Stijepović, and Đorđe Stakić. 2018. "GLAM projekat: galerije, biblioteke, arhivi i muzeji na internetu." In *Zbornik Radova – Asocijacija informacijskih stručnjaka, bibliotekara, arhivista i muzeologa (BAM),* 10:191–193. https://www.ceeol.com/search/article-detail?id=846203.

Popović, Aleksandra, Milica Ševkušić, and Đorđe Stakić. 2015. "Biblioteke i Vikipedija zajedno na vebu: slobodno znanje za sve." *Digitalna humanistika: tematski zbornik u dve knjige, knj. 1,* 151–161. https://doi.org/10.18485/dh.2015.1.ch12.

Popović, Aleksandra, Dragana Stolić, and Dejana Kavaja Stanišić. 2019. "Doprinos Univerzitetske biblioteke "Svetozar Marković" stalnom stručnom usavršavanju zaposlenih u bibliotečko-informacionoj delatnosti." *Čitalište* 18 (34): 32–39. https://doi.org/10.19090/cit.2019.34.32-39.

Ratković, Nebojša, and Ivana Madžarević. 2021. "Women's Participation in Wikipedia: Cross-Border Balkan Perspective." *Área Abierta* 21 (2): 237–253. https://doi.org/10.5209/arab.72763.

Singh, Harshdeep, Robert West, and Giovanni Colavizza. 2021. "Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia." *Quantitative Science Studies* 2 (1): 1–19. https://doi.org/10.1162/qss_a_00105.

Snyder, Johnny. 2013. "Wikipedia: Librarians' perspectives on its use as a reference source." *Reference and User Services Quarterly* 53 (2): 155–163. https://doi.org/10.5860/rusq.53n2.155.

Stakić, Đorđe. 2009. "Wiki technology: Origin, development and importance." *Infotheca* 10 (1-2): 61a–69a. http://infoteka.bg.ac.rs/index.php/en/archives/2009/infoteka-10-1-2-2009-69-78.

Stakić, Đorđe, Marija Tasić, Marko Stanković, and Milena Bogdanović. 2021. "Students' Attitudes Towards the Use of Wikipedia: A Teaching Tool and a Way to Modernize Teaching." *Área Abierta* 21 (2): 309–325. https://doi.org/10.5209/arab.72760.

Vetter, Matthew A. 2014. "Archive 2.0: What composition students and academic libraries can gain from digital-collaborative pedagogies." *Composition Studies* 42 (1): 35–53. https://www.jstor.org/stable/compstud.42.1.0035.

Open Science on Wikipedia: Libraries' Activities in Serbia

Aleksandra Popovic
popovic@unilib.bg.ac.rs
Sanja Antonic
Oja Krinulovic
Djordje Stakic

Open Science

Open science should make scientific information available. There is a tendency that scientific results are published immediately after the research. The application of open science has already shown that in that way science develops faster and raises the level of education of people. Guide to Open Science was published in 2014. Open science makes general science more transparent and accessible to a bigger number of researchers. Open science enables faster measurement of bibliometric indicators which influence both research and researchers (citation, h-index, Journal Impact Factor).

Wiki Librarian

In 2015 the University Library "Svetozar Marković" in Belgrade, Serbia, started a very successful project Wiki Librarian in cooperation with the Faculty of Mathematics of the University of Belgrade and Wikimedia Serbia and the project has continued in 2016. The aim of the project is writing articles on Wikipedia in Serbian and editing pages on other Wiki Projects. Librarians are experts in finding valid and relevant sources of information. They have access to reference collections, commercial and OA e-journals, e-books from rich library collections and printed publications. Many libraries in Serbia joined the project whose aim is to put articles on various topics in OA and to make them available to everyone via Wikipedia.

Wiki Marathon

Wiki Marathon is an event organized within Wiki Projects. Its topic is defined in advance and all participants either write or edit articles on that specific topic on Wikipedia. It can last a day or several days. Both trained Wikipedia editors and beginners can participate.
Wiki Marathon Open Science was organized within the project and articles on the topic were created on Wikipedia in Serbian. They are connected with the articles on Open Access Movement and Open Education, OA software, digital repositories and valuable sources of information. New articles were written also about the specific following topics: Budapest Open Access Initiative, Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Cape Town Open Education Declaration and Bethesda Statement on Open Access Publishing.

Wikipedia

"Wikipedia is a free encyclopedia which everyone can edit". Wikipedia is used as one of the first sources of information. In order for an article on Wikipedia to be valid it has to be verifiable. That means that it must have appropriate references and additional literature. Wikipedia in Serbian was established in 2003. The title page was created in February 2003, and the first articles were published in September. Since then the number of articles has been increasing, and there is a trend in article extension and quality improvement. At the moment there are 291 languages on Wikipedia.

References

Guide to Open Science on link http://digitheadslabnotebook.blogspot.rs/2014/01/guide-to-open-science.html
Simcha Jong, Kremena Slavova (2014). *When publications lead to products: The open science conundrum in new product development.* Research Policy, Vol. 43, Iss. 4, Pages 645-654

design by Aleksandar Milosevic

22nd annual
CONFERENCE ON PROFESSIONAL INFORMATION RESOURCES

Figure 1. Poster from the conference INFORUM 2016

Figure 2. Poster from the conference INFORUM 2017

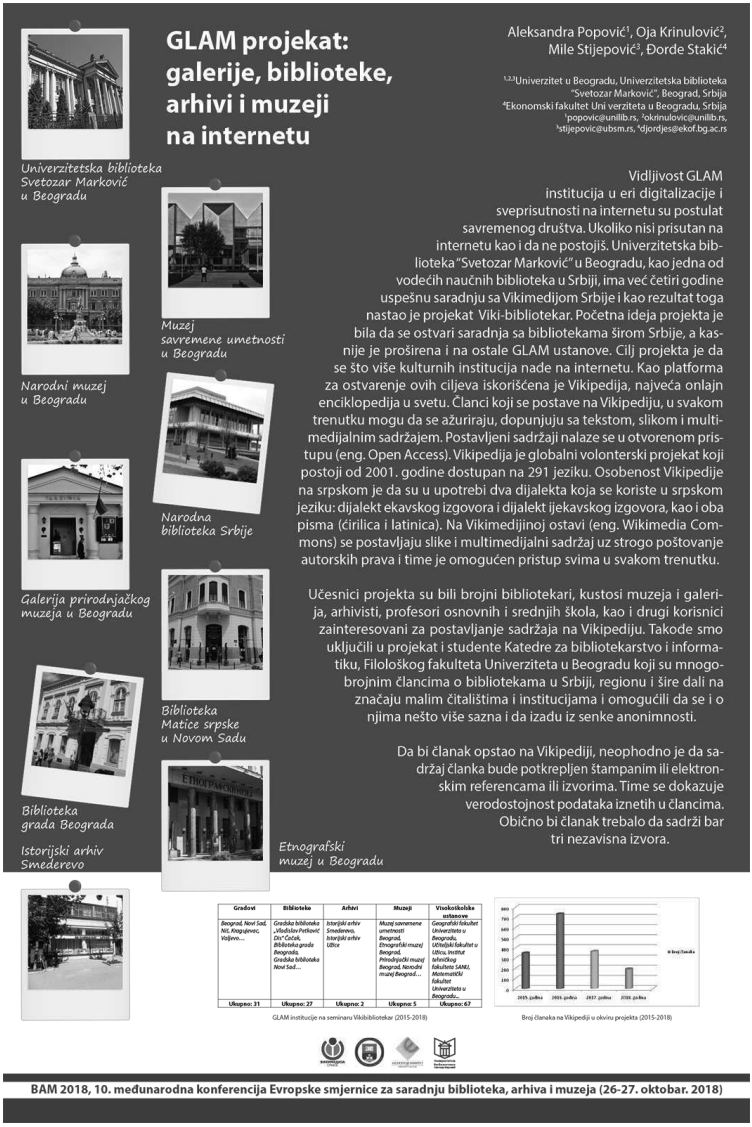


Figure 3. Poster from the 10th International Conference of the European Guidelines for Cooperation of Libraries, Archives and Museums (Sarajevo 2018)

Social Media and Its Role in Amplifying a Certain Idea of Beauty

UDC 316.62: 004.738.5
DOI 10.18485/infotheca.2021.21.1.4

Adeeba Siddiqui (GL-7416)
zarasiddiqui86@gmail.com
Aligarh Muslim University
Department of English
Aligarh, India

ABSTRACT: The use of social media is gaining momentum day by day. It has become extremely popular and has set its foot everywhere in society. Social media has influenced people in a lot of different ways but the most prominent one is the standards of beauty. The paper analyses the relationship between social media and its effect on people's views about their bodies in relation to what they see on social media. And also, how this influence of social media has amplified the use of unrealistic beautification applications and filters among people, especially teenagers and young adults. These applications and filters are photo-editing tools which allow the user to alter their images and make them prettier. These morphed images are unrealistic and can sometimes lead to low self-esteem in young women. A small survey was conducted in which twenty-six young women have participated and shared their views regarding social media experiences. The aim is to study how social media has altered their views about a female body and beauty standards.
KEYWORDS: social media, perception of beauty, beautification application.

PAPER SUBMITTED: 16 June 2021
PAPER ACCEPTED: 18 July 2021

1 Introduction: Use of Beautifying Applications and Filters

Social media has its wings all spread in our day-to-day lives. Nowadays, whoever uses a smartphone is inclined to be on some social media platform.

According to a digital global report (Kemp 2021), social media growth has accelerated significantly since the outbreak of Covid-19. The number of users has risen by more than 13% over the last year, with approximately five hundred million fresh users taking the global user aggregate to nearly 4.2 billion by the beginning of the year 2021. The advancement in digital technologies is increasing rapidly. Social media has a huge influence on people nowadays and especially among the young adults. One of the principal features of Instagram and Snapchat is texting as well as image and video sharing. The popularity of these social media platforms is immensely huge and is increasing day by day. A study (Iqbal 2021b) shows that the greatest number of users are of age between 18–24 on Instagram and India is second highest on the list.

Fashion, Skincare/Cosmetics and Health & Fitness (Iqbal 2021a) are in the list of top ten interests of Instagram users. Kim Kardashian West, an American model, is the 7th most followed person on Instagram. Kendall Jenner, Khloe Kardashian and Kourtney Kardashian are among the top twenty most followed models on Instagram. All these women have a similar body type and facial features. Each one of them has a defined jawline and high cheekbones, contoured nose and poutier lips. Even their body shape is quite similar as they have slim arms, legs and waists with enhanced bottoms and breasts. These women have artificially acquired their beauty,[1] they have immensely changed their bodies through plastic surgeries like lip lift surgery, Botox, rhinoplasty, lasered hairline and what not. If one looks at Instagram profiles of fashion influencers in India like Komal Pandey, Diipa Buller Khosla, Kritika Khurana, Aashna Shroff, Pooja Mundhra, Santoshi Shetty, their images have a lot of similarity in terms of poses and make-over. Their lips are poutier, nose contoured, defined jawlines, high cheekbones, pear shaped body figure. These qualities have somehow become intrinsic for being beautiful. It has created a sense of uniformity on social media profiles.

Although beauty is a subjective concept its definition can differ from person to person. Social media can be seen playing an important role in framing a certain body type to qualify for being beautiful. The standard beauty type in our society is defined by the modalities of advertising, marketing, mass image reproduction which normalizes a certain idea of beauty. This idea is again plural but nonetheless in those pluralities there are certain intrinsic qualities. In todays' time, poutier lips, enhanced eyes, defined jawline, contoured nose, high cheekbones, toned body have become intrinsic qualities,

1. See the article by Mehera Bonner "These Before-and-After Pics of the Kardashians Will Blow Your Mind" in *Cosmopolitan* (Dec 9, 2020) Article

especially on social media platforms. So, there are many applications that are picking up on these essentials of beauty which are also being manifested in Snapchat and Instagram filters. People who can't afford expensive make-up and plastic surgeries can alter/edit their images using these applications and filters in order to be a part of the normative of beauty. Many young women can be seen using them to enhance their lips, eyes, complexion and have a more toned face.

The fashion and beauty magazines like *Vogue*, *Elle*, *Allure*, *Femina*, *Cosmopolitan*, etc., have been playing a vital role in defining a certain idea about beauty for ages. Images of celebrities and models displayed in these magazines are admired by the masses and they frame this certain idea of beauty. The images displayed are often photoshopped or edited for the women to look picture perfect. This idea (Lindblad 2020) of photoshopping goes back to more than thirty years, when two brothers, Thomas and John Knoll, developed Photoshop in 1987 and sold the distribution license to Adobe Systems Incorporated in 1988. It was the first photo editing software introduced to the masses. Photoshop developed over the years with many new versions which took digital editing to a next level. Gradually, many simpler applications like Beauty Plus, PicsArt, B612, Snapchat, YouCam, Beauty Booth, Photo Wonder, etc. were released for smartphones.

Technological advancement has been radical over the years. These applications and beauty filters have created a huge difference among people regarding the idea of beauty. People are immensely using them to beautify themselves and to project a simulated self on virtual platforms. Jean Baudrillard, a French philosopher, in his work "Simulacra and Simulation" (Baudrillard 1994) has analyzed the connection among reality, symbols and society. He is of the opinion that the current society has "replaced all reality and meaning with symbols and signs, and that human experience is a simulation of reality." Simulation has been defined as an imitation which substitutes for the real world by Baudrillard. He talks about how contemporary media is responsible for obscuring the distinction between products actually required and products for which the requirement is created by the marketing industry. For the postmodern society, he says that "it is no longer a question of imitation but substituting the signs of the real for real".

This idea of a perceived reality works in the use of photo editing apps and filters which have made it easier to look prettier than your usual self. People use these filters and apps to create a hyperreal image of themselves on social media platforms. The concept of hyperreal is defined as something "more real than real". For instance, film stars in movies look glamorous

and are so believed to be. Their real image is different from their camera image. One example can be Rajnikanth, an Indian actor primarily working for South Indian films who looks older in reality than his film star image on screen. Thus, people have formed a hyperreal image of themselves to appeal to audiences.

These filters and apps can actually do magic to your images. One can apply make-up, lighter the skin tone, thinner the nose, apply fake bigger eye lashes, lipstick, enhance lips, bosoms, bottoms. It has become a popular trend to put up a crystal-clear picture of yourself on social media platforms. The use of these filters and applications has made it easier to shine in images even when you may have just woken up from bed. To illustrate further one of the participants of the survey has actually agreed to let her pictures be used for the study (Table 1). One can clearly make out the differences between the two pictures. The skin is smoother, nose thinner, lips poutier, breasts enhanced. The application used for these changes is Beauty Plus.

Table 1. Left: photo before modification; right: photo after modification.

One interesting thing about these filters and applications is that they are more emancipatory than surgeries. One can remain in their actual skin but alter their face and body using these apps and project a simulated self for virtual masses. This idea is attractive for people posting their pictures only for monetizing their accounts or just for more appreciation from the audience. Many influencers on Instagram use these filters and applications to create a prettier image on virtual platforms which sometimes help them

to gain immense popularity (for instance, a greater number of followers and likes) and eventually more money.

It is quite fascinating as there are many people who are using these filters and apps to fit into that certain idea of beauty to gain more popularity and hence more money. Influencer marketing is growing day by day and brands are compensating more money than ever before. Many Influencers are promoting products of different brands and making huge amounts of money. Brands usually affiliate links or promo code-based commissions for influencers promoting their products. There are different factors on which the earnings depend like quality of content produced, number of followers, profile's engagement, etc. Instagram Marketing Hub is one of the influencers calculators (Schaffer 2021) used to measure the profiles of the influencers considered. This particular calculator focuses on three factors: size of the followers, average "likes" per post and engagement rate. Hence, it can be concluded that filters and apps have gained huge popularity among Instagram users and influencers have helped in a way to amplify its use.

S No.	Followers	Average Likes	Engagement per Post	Estimated Earnings per Post in USD
1.	97,277	2,850	2.95%	$390–$650
2.	12,412	903	7.67%	$84.75–$141.25
3.	8,559	438	5.25%	$79.5– $132.5
4.	24,767	849	3.43%	$153.75–$256.25
5.	69,367	1,136	1.72%	$277.5–$462.5
6.	596,535	21,164	3.58%	$1,782–$2,970

Table 2. Details of the Instagram profiles of six influencers

Table 2 presents details of the Instagram profiles of several Indian influencers that are presented in more details in text that follows. These details have been taken from Instagram Marketing Hub, a website (Geyser 2020).

1. An Indian TV actress, blogger and an Instagram social influencer who creates humorous short reels in which some or the other filter is used to either enhance the background or beautify her face. She has the perfect poutier lips and thinner nose through surgeries. She has around ninety-seven thousand followers and publishes content mostly about fun and humor. According to her Instagram Marketing Hub report, she is earning money somewhere between 29,098.10 and 48,496.83 Indian rupees (INR) per post.

2. A physiotherapist and blogger who also calls herself an Instagram Influencer. She has used filters in almost every reel to brighten up her skin. She promotes different products like vitamins' supplements and skincare products from famous international companies. She has more than twelve thousand followers and is making money between 6,323.4–10,538.73 INR per post.
3. A Fashion Influencer who is making around seventy-nine to one hundred and thirty-two dollars (5,931.53–9,885.89 INR) per post with more than eight thousand followers. Most of her pictures and reels are filtered.
4. The fourth example is of a Instagram video creator and most of her content is about cosmetics especially make-up products. She has a good number of followers and is making between 11,471.36–19,118.94 INR per post.
5. An Instagram influencer has used a good amount of filters in her dance reels. She promotes different products of skin and hair care, jewelry, clothes, etc. She has a good number of followers and is making about 20,704.5–34,507.36 INR per post.
6. An Instagram influencer and model who is also a TikTok queen and calls herself a "Barbie Doll". She has a huge fan following with more than half a million followers on her profile. She is earning between 132,955.91–221,593.18 INR per post which is an enormous amount of money. She looks similar to a barbie in her reels and pictures with the help of make-up and filters. One can easily identify the immense use of filters in her posts. She acts and talks like a barbie with fixed facial features in a robotic manner as if she is the toy-barbie doll.

2 Data Collection and Methodology

The research work is based on the primary source of data collected through sending out questionnaires to young women using social media platforms. The main aim of the questionnaire was to know how many of these women were using beautifying apps and filters for their images. Also, it aimed to know whether the use of these filters and applications created any sense of low self-esteem in them about their real skin and bodies. The questionnaire was sent out randomly mostly via WhatsApp Messenger and Instagram to around thirty women but not everyone was comfortable responding to such questions. Twenty-six have responded and the analysis has been made accordingly.

Questionnaire:

1. What is your name?
2. What is your age?
3. What do you do?
4. What is your perception of beauty?
5. Which mobile phone do you use?
6. Do you follow any model on social media platforms? If yes, name them.
7. Do you use any sort of beautification apps or filters for your pictures?
8. Do you think your filtered pictures are better liked than unfiltered pictures?
9. How has the use of beautification apps and filters changed the perception of beauty?
10. Has their application affected your self-esteem?
11. What do you usually alter about your face or body using these filters or apps?
12. If you had enough money, would you opt for surgery as an option to change any part of your face or body?

The greatest number of responses (18) are from college students ranging between the age group 18–24 years which is, as stated in the beginning of the paper, the greatest number of users on Instagram (Iqbal 2021a). The age group 18–24 includes the transitional period from adolescence to adulthood in a person's life. It is easily visible that women turning into adults are more conscious about their looks than older women. The participants also include a few working women and one homemaker. The survey targeted only women consciously as the popularity of filters and beauty apps is more among women than men. Though lately one can notice younger men are also being attracted to the use of filters and apps.

Graphical representations of the data collected according to the mentioned categories is presented in Figure 1.

3 Analysis of the Data Collected

On the basis of the data collected, it can be said that because of the unrealistic beauty standards these women have felt insecure about their looks which has led them to use filters and beautification. Twenty-four out of twenty-six said that they are using filters or some sort of beautification app either to lighten their skin or give background noise. All these women agreed that their pictures are better liked when posted using filters. These beauty standards have created a certain idea of beauty where women feel that they need

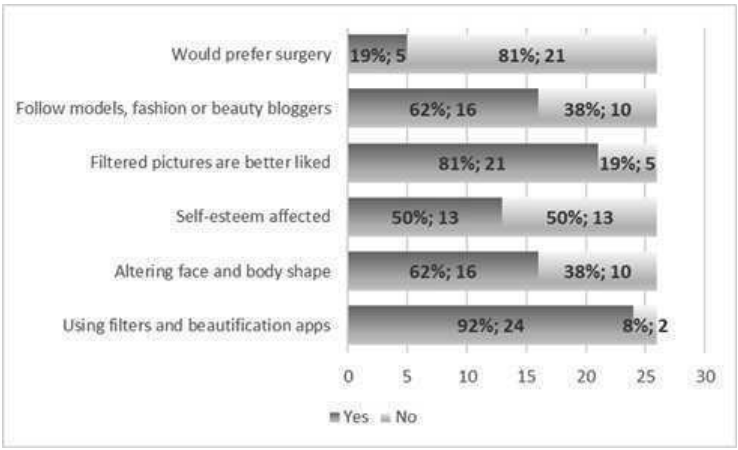


Figure 1. Analysis of answers to questions 6, 7, 8, 10, 11, 12

to alter their bodies according to their needs. Poutier lips are in trend these days and one of my participants is even ready to undergo surgery to enhance her upper lip. Perfect jawlines and high cheek bones are in demand too and most of the women accept that they usually alter their jawlines using these applications and filters.

Sixteen of the women participants follow models and other fashion or beauty bloggers who promote this certain image of being beautiful where they have perfect jawlines, poutier lips, enhanced lips and eyes, high cheekbones with slim waist. These models are mostly plastic but they have created a beauty normative which has become an ideal for most women. Younger women are in awe of such celebrities and models and admire them for their perfect faces and bodies. So, most younger women feel an urge to look like them and take support of filters and applications to alter their image. The fashioned contoured nose is very much in demand. And the five participants who were ready to go for a surgery mostly preferred a rhinoplasty in which one's nose is altered and reconstructed which changes the appearance of the nose.

The questionnaire included a question where the participants' perception about beauty was asked. The majority of participants answered that beauty is something which resides inside of a person. Also, a confident person seems beautiful to many participants but at the same time these women are accepting that they use filters and apps to beautify their images which in itself is a contradictory notion. Also, the patriarchal notion of woman as a beautiful

object has triggered the use of filters and apps. In a patriarchal mindset, the existence of a woman is limited to a man's needs and the idea that a fair wife is a beautiful wife has also forced women to make themselves look appealing to a man. This is an unhealthy mindset which compels younger women who are easily convinced to take support of make-up, filters, beauty apps and the extreme option surgery in order to look pretty in accordance with the normative of beauty.

Many participants expressed that the constant use of social media has led them to compare themselves to the beauty images, and also agreed that use of these Snapchat and Instagram filters have altered their opinions about beauty. One of the participants said that these filters used by so many people belittle those who already lack self-confidence and are uncomfortable in showing their real skins. They give the viewers an illusion of extremely fair and clear skin, a thinner face, thinner nose and what not. These pictures might mess up with people's heads who start comparing these filtered pictures with their real skin. These filters and beautification apps have definitely changed the perception of beauty because when you use filters just make blemishes disappear. You can make anything disappear. You can make skin smoother which many times leads young women to believe that their skin should be flawless and they should not have pores in it. And that obviously creates a toxic environment and it also creates toxic things where women are forced to believe that the person you are looking at the filter is better than the person you are looking at the mirror.

Self-esteem is another important part of one's life. It is a subjective evaluation of one's own worth. The study also involved whether women when comparing themselves to their filtered hyperreal selves felt that it has hurt their self-esteem in any way. Thirteen participants answered that in a way their self-esteem was hurt, especially, when they started comparing their filtered self with their mirror image. The person in the filter is not real but when these women started aspiring to become like them, it did push their self-esteem. Many agreed that extensive use of these filters have now made it impossible for them to post a picture without filters. They do not feel confident in posting a picture without editing anymore. Some even said that they only use a filtered camera rather than a normal camera.

If we go deeper into psychoanalytic theory, this desire to look like your filtered self by opting for surgery reflects on the person's *id*, the impulsive and unconscious mind of a person. According to Freud's psychoanalysis theory (Freud 1990), psyche is framed into three parts: *id*, the primitive and instinctual component responding to basic desires and needs; *super-ego*, the

moral conscience and; *ego*, the realistic part mediating between *id* and *superego*. These apps and filters give an unrealistic image of oneself but the desire to look that way and to fit into the certain idea of beauty just to satisfy their *id* is reflecting upon the extensive effect of the use of these filters and apps.

Id works on the pleasure principle and wants every wishful impulse to be satisfied. Surgery is a radical option which is painful and can have serious consequences. Even after that if one still aspires to be like her filtered self via surgery, it is a problematic issue as one is putting up with something dangerous and uncertain. Many times, these surgeries fail and the outcomes are horrifying. Letting your unconscious mind rule over your rational self and opt surgery as an option is not at all a good idea.

4 Conclusion

With the westernization of India, the beauty standards have changed a lot in the recent years. The idea that "fair is beautiful" has been deeply rooted in Indian society. The matrimonial advertisements have always emphasized on the fairer skin tone of the girl. There is even particular demand for fair-skinned girls mentioned in the advertisements (Gelles 2020). Among all the qualities in a girl, fair skin tone is an intrinsic quality for being called beautiful in a stereo-typical Indian society and also, the most demanded. Colorism is deeply inscribed in Indian mindset and especially in the case of a girl which makes it a gendered bias. The Indian fairness cream market revenue was reportedly worth 30 billion dollars in 2019 and is anticipated to cross 50 billion dollars by the year 2023.[2] This idea is engraved in the minds of Indians that fair skin tone is the most essential quality for a woman to be beautiful. After this, a woman with bigger eyes, long black hair, full red lips and slim waist with wider hips is considered to be the ideal beauty in Modern India (Majidi 2020). These essentials are somewhat different than the new globalized beauty trends, for instance, fairness is no longer a significant quality due to racial equality in the global beauty trends. Also, the idea of thick-black long wavy hair is accustomed to Indian beauty ideals. The global beauty trend of poutier lips is also not a part of the traditional ideal of beauty in India. Discrimination and racism do not only happen between communities

2. See the article by Pia Krishnankutty "Before Fair & Lovely, there was Afghan Snow – all about the fairness creams market in India." in Print (June 26, 2020) Article

but also within communities. Biasness for lighter skin tone exists among many countries across Asia, Africa and South America.[3] According to a study (Sims and Hirudayaraj 2016), women that are dark-skinned in India and surrounding countries relatively have lesser opportunities to become a flight attendant, journalist, receptionist, model, sales associate, actress and other professions which needs interaction with masses because they will be judged as "unattractive". Even among the participants of the survey, six to seven women directly or indirectly admitted that they lighten their complexion or skin tone using filters or beautifying apps. Most of the participants in the survey are either college students or working, hence, it can be stated that level or type of education does not matter when it comes to the use of filters and beautification apps on social media. Also, highly qualified people like doctors can be seen posting dance or music reels during the ongoing Covid-19 pandemic on social media platforms in which they usually have applied filters.

Therefore, it can be concluded that social media plays an important role in defining a certain normative of beauty as all the top models and fashion bloggers look in a similar fashion, dress up in a similar fashion and even post pictures of a similar fashion. These trends have led other women to follow in their footsteps where these women take help of digital apps and filters to imitate them. Nowadays an image of a beautiful face and body is created which has defined certain essentials of being beautiful. It is mainly the enhanced lips and eyes, smoother skin, high cheekbones, defined jawline, slim waist, enhanced bottoms, thinner nose, etc. Many women have been influenced due to the extensive involvement of social media in todays' time. Models and fashion bloggers have defined these standards of beauty which create an urge among other younger women to imitate. And many influencers are using these filters and apps for monetary reasons to gain more attraction and hence money. So, many women are using filters and applications which help them define their jawlines, poutier their lips, smoother their skins, slim their waists, etc. It has certainly in some cases lowered the self-esteem of women where they are unable to feel satisfied with their real skins and bodies. Also, it has created an illusion in them as they aspire to be like their filtered images. This leads to a sense of body dissatisfaction and overall lower self-confidence in many cases. It has become almost impossible for these women to post pictures without filters which is a negative effect. And those who

3. See the article by Meera Estrada "Commentary: Shadeism is the Dark Side of Discrimination We Ignore." in Global News (Posted May 24, 2019), Article

aspire to their filtered self to an extent that they may opt surgery as an option reflects on the mental impact of these filters and apps.

This study has mainly focused on women of young age. Further advancement in the study can be brought about by including young men to see whether they are under the similar urge to enhance their virtual image on social media platforms and also, to women of older generations. The constant pressure from social media to fit into the normative of beauty has affected people's perception of beauty. Also, this extreme urge may lead people to opt for surgery which reflects upon a person's id i.e., the impulsive and unconscious mind. Widening the psychological scope of the study, this aspect of a person's id overruling the rational self can be further explored.

Acknowledgment

This paper was written during author's M.A. studies in English Literature at Aligarh Muslim University, India.

References

Baudrillard, Jean. 1994. *Simulacra and simulation.* University of Michigan press.

Freud, Sigmund. 1990. *The Ego and the Id.* Edited by James Strachey. W. W. Norton & Company.

Gelles, Rebecca. 2020. "Fair and Lovely: Standards of Beauty, Globalization, and the Modern Indian Woman, Independent Study Project (ISP) Collection." Accessed June 22, 2020. http://thelionandthehunter.org/norms-of-beauty-in-india-fair-is-beautiful-a-legacy-of-colonialism-and-globalization/.

Geyser, Werner. 2020. "Influencer Marketing, in Instagram Marketing Hub." Accessed October 29, 2020. https://influencermarketinghub.com/instagram-money-calculator/.

Iqbal, Mansoor. 2021a. "Instagram Revenue and Usage Statistics, Business of App." Accessed May 24, 2021. https://www.businessofapps.com/data/instagram-statistics/.

Iqbal, Mansoor. 2021b. "Snapchat Revenue and Usage Statistics, Business of App." Accessed May 6, 2021. https://www.businessofapps.com/data/snapchat-statistics/.

Kemp, Simon. 2021. "Digital 2021: Global Overview Report." Accessed January 27, 2021. https://datareportal.com/reports/digital-2021-global-overview-report#:~:text=Social%5C%20media%5C%20user%5C%20%20numbers%5C%20increased%20by%5C%20the%5C%20state%5C%20of%5C%202021.

Lindblad, Mason. 2020. "The History of Photoshop – Photoshop Through the Years, Filtergrade." Accessed September 22, 2020. https://filtergrade.com/history-of-photoshop-through-the-years/.

Majidi, Khesraw. 2020. "Norms of Beauty in India Fair is Beautiful: A legacy of Colonialism and Globalization, The Lion and the Hunter." https://digitalcollections.sit.edu/isp_collection/1145/.

Schaffer, Neal. 2021. "How Much Should I Pay an Influencer? Testing Out 9 Instagram Influencers Calculators, in Influencer Marketing." Accessed June 6, 2021. https://nealschaffer.com/instagram-influencer-calculators/.

Sims, Cynthia, and Malar Hirudayaraj. 2016. "The impact of colorism on the career aspirations and career opportunities of women in India." *Advances in Developing Human Resources* 18 (1): 38–53.

Infotheca (Q25460443) in Wikidata

UDC 004.62: [030:004.738.5
DOI 10.18485/infotheca.2021.21.1.5

ABSTRACT: Wikidata is a Wikimedia Foundation knowledge base, a common source of various kinds of data used not only by other Wikimedia projects, but also increasingly by numerous semantic web applications. In this paper, we will present an example of integration of Wikidata with digital libraries and external systems, as well as the potential for speeding up the process of data preparation and entry using the articles published in *Infotheca*, *Journal for Digital Humanities* as an example.

KEYWORDS: Semantic Web, Open Linked Data, Wikidata, Infotheca, journal metadata.

PAPER SUBMITTED: 24 June 2021
PAPER ACCEPTED: 16 July 2021

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

Lazar Davidović
lazarmdavidovic@gmail.com
University of Belgrade
Belgrade, Serbia

1 Introduction

Wikidata[1] is a Wikimedia Foundation knowledge base, a common source of various kinds of data, both concrete and abstract. The stored data can be used by other Wikimedia projects, such as Wikipedia and the wider community alike, for different purposes. This contributes to extending the boundary from machine readable to machine comprehensible data on the web. In this paper, we will present an example of integration of Wikidata with digital libraries and external systems, as well as the potential for speeding up the process of data preparation and entry using the articles published in *Infotheca*, *Journal for Digital Humanities* as an example.

Semantic web is an extension of the existing web where information is given precise meaning allowing better collaboration between computers and their users. The open and partially structured nature of the resources whose development was organized by Wikimedia provided the basis for the creation

1. Wikidata

of many machine readable resources, like DBPedia,[2] for example, relying on the standardized languages of the semantic web. The concept of the semantic web and open linked data technologies expand the traditional web by using a standard markup language and similar processing tools, where RDF (Resource Description Framework) plays a significant role and makes more efficient information retrieval solutions possible (Shah et al. 2002). In order for the semantic web to act the part, computers should have access to structured collections of information and be able to set out defined rules of automated management. Wikidata actually fits the trends of information technology development that extend the boundary from machine readable to machine comprehensible data on the web.

The Scholia project[3] (Nielsen, Mietchen, and Willighagen 2017) is one of the first comprehensive endeavours of its kind aimed at representing bibliographical data, scholarly profiles of authors and institutions using Wikidata. The results of this particular project and the availability of the Infotheca articles in different digital formats provided the inspiration for the "Wikification" of the articles published in the Infotheca journal. Having seen the content of the web pages scholiaEvent[4] and scholiaTopic,[5] a similar project was launched with the goal of creating linked (RDF) data about authors and scholarly articles based metadata and adding links to Wikidata to the journal articles, as well as showing the co-authorship graph as an interactive page on the website of the Biblisha digital library[6] (Stanković et al. 2015). The implementation is a case study that can be further extended to other use cases, such as conferences and digital libraries. The Scholia tool is being developed as part of a lager initiative, WikiCite,[7] aiming to index bibliographic data in Wikidata on the resources that can be used to corroborate the claims made in Wikidata, Wikipedia or elsewhere. At the time when we are inundated by false information on the web, proper corroboration of information by relevant sources certainly plays a key role. Since we wanted to automatize the process of preparing and entering information, we looked

2. DBPedia

3. Scholia, Scholia at Wikidata

4. An overview of past and present conferences with an organizer containing information about article submission deadlines ScholiaEvent

5. An overview of scholarly and professional articles, as well as their authors and topics appearing together, grouped by thematic entities: Wikipedia, machine learning, biology, food and the like. scholiaTopic

6. Biblisha

7. WikiCite

into different solutions and ended up using two of them, namely, OpenRefine[8] and QuickStatements[9] that will be discussed further in the sections to follow.

The collaboration of Wikimedia Serbia[10] with the University of Belgrade is a longstanding one (Stakić 2009). The University Library Svetozar Marković together with the Faculty of Mathematics of the University of Belgrade and Wikimedia Serbia launched the (Wiki-Librarian) project in 2015 with the idea of making as much quality content as possible available on Wikipedia (Popović, Ševkušić, and Stakić 2015). Wikiadata, as an open data network was used by Andonovski (Андоновски 2020) to describe language resources, namely, novels forming part of the Serbian-German literary corpus (Andonovski, Šandrih, and Kitanović 2019). For a number of years now, students at the Faculty of Mining and Geology have been undergoing training to enter data into and use the Wikidata[11] database, while the Intelligent Systems doctoral course features the subjects Knowledge Representation and Semantic Web that explore the potential for application of open data. As part of the "Distant Reading for European Literary History"[12] се ради на уносу метаподатака о српским романима из корпуса *srpELTeC*[13] COST Action CA16204 (2017-2021) metadata about Serbian novels included in the srpELTEC corpus is being entered into the knowledge base (Krstev et al. 2019) and Wikidata linked to various applications, one of which is Aurora.[14] Members of JeRTeh Language Resources and Technologies Society[15] too contributed to the results presented in this article.

8. OpenRefine (formerly Google Refine) is a tool for working with messy data: cleaning it; transforming it from one format into another, together with extending it with external data via web services. OpenRefine

9. The tool for editing Wikidata items: adding and removing statements, labels, descriptions, etc. QuickStatements

10. Wikimedia

11. Input data to Wikidata and their use

12. One of the most important aims of this action is preparing a multilingual corpus (titled European Literary Text Collection - ELTeC) which, when fully complete, will feature a hundred novels from each participating country first published in the period 1840-1920.

13. srpELTeC

14. Aurora

15. JePTex

2 Wikidata

Wikidata is a knowledge base whose purpose is to be a common source of certain kinds of data (for example, the population of a country, place of birth, date of founding) used by other Wikimedia projects, such as Wikipedia. In that sense, it is similar to Wikimedia storage where media files accessed from other Wikimedia projects are stored. Wikidata is oriented towards documents, focused on items representing topics, concepts or objects. Every item has been assigned a persistent identifier, a positive integer with an upper-case Q as a prefix, known as *QID*. This makes translation of the basic information necessary for recognizing a topic covered by an item without favouring any language whatsoever, the aim being to ensure the uniqueness of meaning of a particular concept.

These are some of the examples of items: places (Novi Sad: Q55630, London: Q84, Zvezdara (Belgrade): Q12645852), people (Đorđe Balašević: Q342045, Tim Berners-Lee: Q80, Hedy Lamarr: Q49034), events (First Serbian Uprising: Q368689, concert: Q182832, marathon: Q40244), objects (chair: Q15026, glass: Q81727, frying pan: Q127666)), concepts (joy: Q935526, fear: Q44619, concept: Q151885), literary works (*Gorski vijenac*: Q1192476, *Don Quixote*: Q480, *Game of Thrones* (books): Q1751870), films (*Lepota poroka*: Q4239792, *Hair* (film): Q757156), TV series (*Game of Thrones*: Q23572, *'Allo 'Allo!*: Q425628), ballet (*Don Quixote* (ballet): Q1239463)... The concepts behind items should be unique, but as it happens, there can exist two items under the same name, Nikola Tesla (Q9036) refers to the famous scientist, while Nikola Tesla (Q2732597) refers to a housing project (Q486972) in Niška Banja (Q954986) named after him. It is recommended that in the case of polysemous entities, like the above-mentioned *Don Quixote* (ballet) or *Game of Thrones* (books) an additional explanation should be given in the parenthesis. An item is, thus, linked to a unique identifier (QID), the identifier is, in turn, linked to the item's corresponding title and description, so as to remove any ambiguity.

An identifier of a data item (QID) can, in addition to being linked to a title and a description, have a number of aliases and statements (claims, expressions) representing its properties and values. A statement is an ordered triple (item, property, value), where item (Q) is any topic (person, object, place, concept), item (P) is property . The relation[16] or a characteristic relevant to an item can be, for example: hair colour (P1884) for people,

16. In mathematics, if an ordered pair (x, y) is the relation ρ then the element x is has established a relation to the element y and it is written as a triple: $x\rho y$.

The University of Belgrade is a public university and a member of the European University Association.	
The University of Belgrade was established in 1808 by Dositej Obradović.	
Q240631 P31 Q875538. Q240631 P463 Q868940. Q240631 P112 Q347659. Q240631 P571 "1808".	Q240631 P31 Q875538; P463 Q868940; P112 Q347659; P571 "1808".
Dositej Obradović was born on February 17th, 1742 in Čakovo and died on April 7th, 1811. He was a linguist, poet, writer and philosopher.	
Q347659 P569 "17 February 1742". Q347659 P19 Q325736. Q347659 P570 "7 April 1811". Q347659 P106 Q14467526. Q347659 P106 Q49757. Q347659 P106 Q36180. Q347659 P106 Q4964182.	Q347659 P569 „17 February 1742"; P19 Q325736; P570 „7 April 1811"; P106 Q14467526; P106 Q49757; P106 Q36180; P106 Q36180.

Table 1. Examples of Wikidata items

publisher (P123) for published works, founding (P571) for organizations and the like. The value of an item cam be a literal itself, that is, a character string (for instance: the length of the Danube is 2860 km) or a reference to some other item (the capital of Serbia is Belgrade, for example). An item can be described by a string of statements, each of which provides a fact or a piece of data about the item. Table 1 shows several examples of natural language sentences and the encoding of this information in Wikipedia, represented as triples of subject, predicate and object (left), and in shortened notation (right).

In the above example, the second column of the table features triples i.e. sentences following the subject-predicate-object pattern. More precisely, we could say that these are RDF triples, where RDF is an abbreviation standing for web Resource Description Framework (Q54872). The sentences end in a period. The third column shows abridged notation doing away with the repetition of the subject, so that the punctuation mark "." indicates that the predicate that follows refers to the same subject.

Similarly, the items in Wikidata represent relations as triples, so that the relation Tesla-way of life-vegetarianism is encoded as Q9036 P1576 Q83364.

An important characteristic of Wikidata is that it has two facets. One is intended for people and the other one for machines, which enables numerous applications in the domain of natural language processing. Let's mention some of them: text classification, indexing, text analysis, summarizing, normalizing, linking, etc. Another important feature is multilingual support, making it possible to link each item to a label in any language registered in Wikimedia resources, which, in turn, opens up the possibilities for numerous applications, from automatic translation and classification of multilingual documents, to analysing web and social media content.

3 Automated Wikidata entry

The entry of individual pieces of data is often a time-consuming task, but it can be sped up in the situations where data already exists and is stored in different digital formats. With proper prior preparation, it can be entered in Wikidata semiautomatically. Therefore, the basic idea was to speed up the entry of data about the results of the research in the domain of digital humanities in Serbia, as well as about old Serbian novels, so as to increase the visibility of both the Serbian language, our cultural heritage and the results of the research in Serbia and certainly pave the way for many other data sets.

In order for data entry automatization to be possible, it was necessary to make the first step consisting of collecting and preparing data. The second step refers to the choice of Wikidata labels that will be used to identify the predicate and creating a data entry outline. The outline defines the linking of a value to an item, namely, a subject via predicate as a mediator.

Although the ultimate aim was the entry of data about articles, entering information about their authors was an indispensable step and a prerequisite for further work. After data entry has been completed, SPARQL [17] queries for different views were created, using Wikidata integrated technologies too for visualizing the results.

Every Infotheca journal article from the Biblisha bilingual digital library has been linked to the corresponding Wikidata entry, so that not only each individual, particular Wikidata view can be accessed directly, but also that some of the useful visual representations can be integrated within the application itself. Figure 1 features an example of a pattern in Biblisha showing the three top-ranked articles in the collection and illustrating the fact that behind every article title there is a link to a Wikidata resource path.

17. [18]

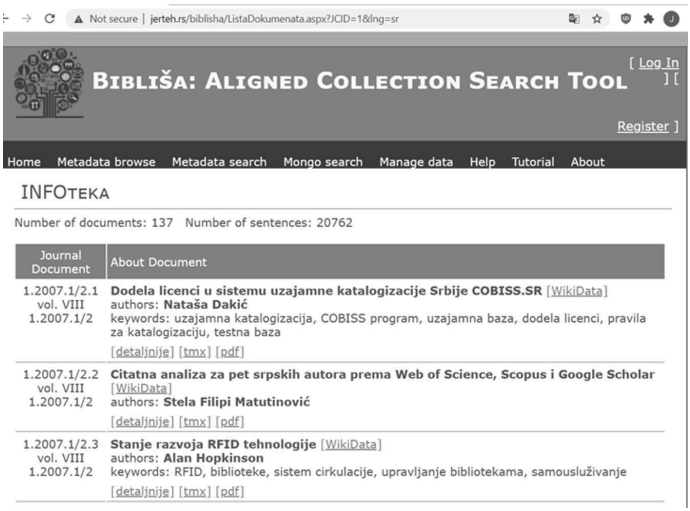


Figure 1. Biblisha digital library panel showing an overview of the metadata about Infotheca.

The information about the first article on the list would be translated into the language of Wikidata in the following way:

„Додела лиценци у систему узајамне каталогизације Србије COBISS.SR“ (Q98785010)
instance of (P21) academic journal article (Q18918145);
author P(50) Наташа Дакић (Q99281474).

We can see that the article is represented by the identifier Q98785010, that it is an instance (P21) of the class of scholarly articles (Q18918145) and that it has the author Q99281474.

The above-mentioned tool OpenRefine initially developed by Google, as well as the QuickStatements tool developed by the Wikidata team member, Magnus Manske are often implemented together and can be said to complement each other. QuickStatements uses textual TSV or CSV formats efficiently generated by the OpenRefine tool. The difference between these two tools is in the granularity of transactions, since OpenRefine inputs the changes in a single step, so resolving data entry errors can lead to data duplicates, while QuickStatements inputs each item individually, allowing better monitoring of the entire process. The examples of good practice indicate that OpenRefine is used for preparing data entry in the Wikidata database, while

the actual entry of RDF triples is performed by using the QuickStatements tool.

Preparing data in the form of a CSV file is certainly the first step, followed by creating an OpenRefine project and loading the prepared data. What comes next is recognition of the existing Wikidata items – an indispensable step enabling the linking of file content to identifiers (QID) of the existing items and entry of new ones, if they do not already exist. In this phase, manual checking and possibly making changes are necessary. Creating a dataset schema defines the predicates that will link subjects to objects in RDF triples and it is a very important step. Here are some characteristic examples with comments:

- Title (P1476), in English and Serbian (both scripts, Cyrillic and Latin, for the sake of search);
- Main subject of the creative work (P921), key words, where the existing ones are linked and new ones are added as instances with labels in Serbian and English;
- Publisher (P123);
- Language of the work or name (P407);
- Publication date (P577), represented by year only;
- Published in (P1433) Infotheca (Q25460443);
- Licence (P275);
- Full text available at (P953).

In view of the fact that properties are added more and more rarely, it is recommended to perform a search of similar properties and properties of similar resources before the decision for them to be added is made. Moreover, special attention should be paid to the limitations in the domain of properties that can be seen in the suggestions provided when entering data. Joint work of distributed users unavoidably sometimes leads to Wikidata duplicates. The solution to this situation is making use of the option to merge or eliminate duplicates. Additional information about it is available at this page.

After the initial data entry, the input of data into the database continued after each newly published issue of Infotheca. As a result, 38 Infotheca articles are now made available. An HTML integrating Wikidata Query Service with Biblisha was created. The queries retrieving tables of the latest published articles, frequency of the keywords in articles, pictures of authors, author profile table, co-authorship graph, distribution of authors by sex, etc. were written.The information about authors consists of basic data that

should definitely be enriched with new content in the forthcoming period, including data about the institutions where they work, research interests, references to authoritative research databases and the like. By way of example, here is a simple query, available at this site, showing a list of Infotheca articles, issue, volume and publication date.

```
SELECT ?paper ?paperLabel ?vol ?publication ?publication_date
WHERE {
?paper wdt:P1433 wd:Q25460443;
       wdt:P577 ?publication_date;
       wdt:P478 ?vol;
SERVICE wikibase:label {bd:serviceParam
                        wikibase:language "en,sr".}
}
ORDER BY DESC(?vol )
```

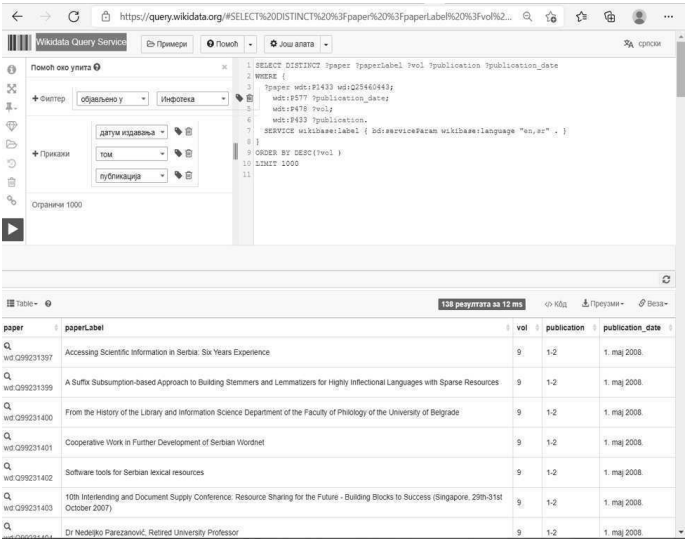


Figure 2. *Wikidata Query Service* query interface

Figure 2 provides an illustration of the above query in the Wikidata Query Service user interface, where the upper part of the panel is used for

making queries, while the results are shown in the lower part. The view (table, graphical representation, grid, timeline, chart, map etc.) can be chosen depending on query type.

Figure 1 features part of a co-authorship graph[19] pulling data from Wikidata via Wikidata Query Service.

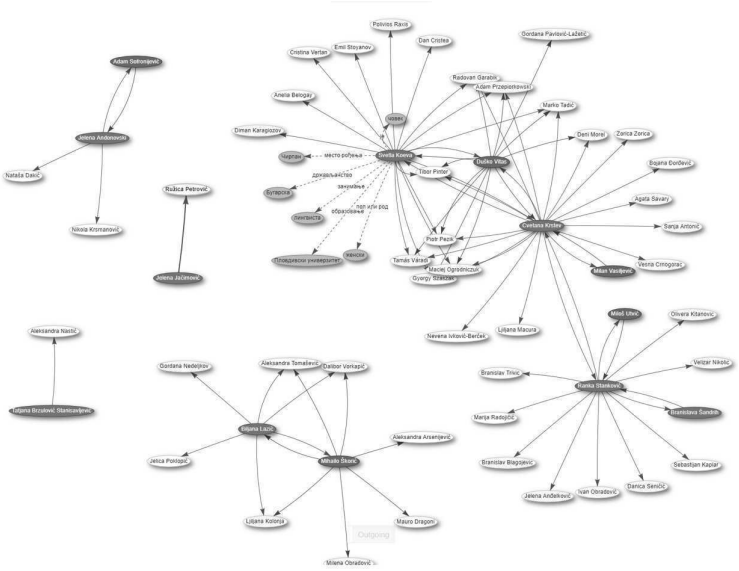


Figure 3. *Infotheca* articles co-authorship graph

4 Conclusion and future plans

Positive experiences of working with Infotheca Wikidata were drawn upon when entering in Wikidata the data on the novels and their authors belonging to the ELTeC multilingual collection (European Literary Text Collection) one subcollection of which will consist of a hundred Serbian novels from the period 1840-1920 developed as part of the Distant Reading for European Literary History Cost Action: CA16204 by members of the JeRTeh society,

19. The co-authorship graph is available, where the SPARQL query itself can be accessed or the view changed to table, timeline, graph and the like.

led by Cvetana Krstev and Ranka Stanković. A set of metadata about the novels digitized up to the present and prepared to fit the requirements of the action was built. The work on Wikidata is seen as a continued activity where special attention will be paid to linked open data in the domain of linguistics – LLOD and its application. We must certainly be aware of the problems and limitations related to Wikidata and other kinds of linked open data, so as to be able to look into the ways of overcoming or at least mitigating them.

Wikidata is a truly immense knowledge base that is 1) available to everyone – for reading information, making queries, editing and improving it; 2) Open – multiple use is available under the Creative Commons CC0 licence granting complete freedom to use data; 3) multilingual – entities can be named and described in any natural language. These three key features are the main driving forces for the many applications, which, we believe, will inspire the wiki community further to devote more attention to this resource. The so-called small languages, including Serbian, should make use of all the possibilities to find their place in the digital space. Thus, the activities carried out as part of this and similar projects and initiatives are a humble attempt to contribute to the preservation of the Serbian language in the digital age.

References

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

Krstev, Cvetana, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. "Analysis of the first Serbian Literature Corpus of the Late 19th and Early 20th century with the TXM platform." In *Book of abstracts of DH_ BUDAPEST_ 2019,* 36–37.

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

Popović, Aleksandra, Milica Ševkušić, and Đorđe Stakić. 2015. "Biblioteke i Vikipedija zajedno na vebu: slobodno znanje za sve." *Digitalna humanistika: tematski zbornik u dve knjige, knj. 1,* 151–161.

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

Stakić, Đorđe. 2009. "Wiki Technology - Origin - Development and Importance." *INFOtheca-Journal of Informatics & Librarianship* 10 (1-2): 69–78.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. "A bilingual digital library for academic and entrepreneurial knowledge management." In *Proceeding of 10th International Forum on Knowledge Asset Dynamics-IFKAD,* 1764–1777.

Андоновски, Јелена. 2020. "Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса." PhD diss., Универзитет у Београду, Филолошки факултет, јануар.

Multimedia Project *The Two of Them*

UDC 004.55:378.147]:02(497.11)
DOI 10.18485/infotheca.2021.21.1.6

ABSTRACT: The topic of the multimedia project *The two of them* is the presentation of novels by Serbian writers in which the characters are having romantic relationships. The final year students of the Department of Librarianship and Information Science at the Faculty of Philology, University of Belgrade participated in the project. The project was realized within the subject Multimedia documents of the academic year 2019/20. Each of the students created a web page containing additional information about the writer and the novel for one chosen novel, using HTML language. In addition, a website containing all the papers in one place was created, as well as a list of students who participated in the preparation of the document.
KEYWORDS: multimedia documents, literature, Serbian novels, Serbian writers, love novels, literature awards.

PAPER SUBMITTED: 4 June 2021
PAPER ACCEPTED: 30 August 2021

Katarina Glavonjić
katarina.26.g@gmail.com

Magdalena Lukić
magdalenalukic383@gmail.com

Jovan Janković
dzovybrat@gmail.com

Stefan Joksimović
joksimovic.calisthenics
@gmail.com

University of Belgrade
Faculty of Philology
Belgrade, Serbia

1 Introduction — About literature and reading

Literature uses language the way painting uses colors or music tones, but language is not only a natural material, but an extremely complex spiritual creation (Solar 2005). Literary works and reading are an important element of human life. "Reading has often been compared to traveling (. . .). As a traveler sees and experiences more than the one who has never left his own doorstep, the reader discovers unknown worlds and other people's lives. (. . .) The worlds of poetic imagination have no limits in space or time and their number is constantly increasing." (Тартаља 1998)

Today, books and reading are increasingly exposed to the competition of new means of acoustic, visual and multimedia communication. So there are

many novels that have their theatrical adaptations or are screened. In the multimedia project *The two of them* novels in a theatrical adaptation can be found along with the screen adapted novels.

2 About the novel as a literary genre

The novel is the most popular and most published literary genre today. From the 13th century, a Chivalric romance appeared, and the name *le roman breton* referred to works (in prose and verse) of Celtic and Breton origin in which the love and warrior adventures of Christian knights were described (Живковић 1992). Novels were considered fictional love stories. Ancient and chivalric novels *owe* the most to old epic poetry and fairy tales, while the newer European novel relies on the tradition of humorous stories, anecdotes and short stories (Тартаља 1998). In the 18th century, the novel became the most represented literary genre. Over time, it became a story about the private world and about individual personalities, in a more realistic and intimate manner. An individual's reaction to the action of the outside world is a key part of developing a plot in a novel. From the 18th to the 20th century, the characterization of individual characters in the novel became more and more important, generalizing the important characteristics of one environment and epoch. The novel depicts the life of a society through the history of the destiny of individuals while the modern novelist has become increasingly interested in man himself (Живковић 1992).

2.1 Types of novels

There are many different types and kinds of novels. We can classify them according to various criteria. They are often classified according to the topic (historical, adventurous, romance, social, psychological, etc.), the method of topic analysis (humorous, satirical, sentimental, etc.), the form of the novel (travel, novel-essay, novel in sequels, etc.), the direction and epoch of their origin (medieval, postmodernist, etc.). Novels can also be classified according to the elements that dominate the novel itself. Thus, a novel can also be: a novel of a character, a family novel (developed from a novel of character which generally follows the members of one family), a novel of space, a novel of time, as well as a novel of time and space, and so on. (Ђорђевић and Лучић 2008)

3 Briefly about Serbian literature and prominent writers from the 19th to the 21st century

The emergence and development of Serbian literature was influenced by many Serbian and foreign writers. We will list only some Serbian writers who are important for the development of Serbian literature, and most of whom are part of this multimedia project.

Of all the modern literary genres, the novel appeared the earliest in our literature. Translations first appeared, while the original novels appeared at the beginning of the 19th century, with Atanasije Stojković (1773–1832) and Milovan Vidaković (1780–1841). Atanasije Stojković wrote the didactic-idyllic novel *Aristid i Natalija* (1801), which represents the first appearance of this type of literature in our language. Jakov Ignjatović (1822–1889) is considered to be the creator and one of the most important representatives of the Serbian realistic novel in the 19th century. Among the important novelists of the 19th century, we should also mention Svetolik Ranković (1863–1899) — the creator of the Serbian psychological novel. He left behind several works, among which is a significant novel — *Seoska učiteljica* (1899), which is one of the first novels depicting a woman as the main character. At the end of the 19th century, in Serbian realistic prose, a lot was written about the village — about the decay of the village or its idyllic images, but also about customs and relations between people.

A great turning point, both in European and Serbian literature, occurred at the end of the 19th and the beginning of the 20th century. At that time, Serbian literature acquired all the basic features of modern national literature. Serbian literature started making its way out of the former *rural* realism, at the time. Borisav Stanković, Petar Kočić, Milutin Uskoković, Veljko Petrović, Isidora Sekulić and others are among the prose writers who went towards the modernization of prose expression. Borisav Stanković (1876–1927) is the most important narrator in the modern phase of realism. The development of the Serbian modern novel and short story begins with him, while his novel *Nečista krv* (1910) is one of the best novels in Serbian literature.

The twentieth century affirmed women's creativity in literature more than ever before. Along with Isidora Sekulić (1877–1958), a significant role in literature had Jelena Dimitrijević (1862–1945) and Milica Janković (1881–1945). It can be said that the novel *Đakon Bogorodičine crkve* (1919) by Isidora Sekulić and the novel *Nove* (1912) by Jelena Dimitrijević are among the best novels written by women at the beginning of the 20th century in Serbian literature. Milica Jakovljević (1887–1952), better known under the

pseudonym Mir-Jam, is a writer whose love novels and short stories marked domestic literature between the two world wars. She was the most popular writer in the Kingdom of Yugoslavia and her works were the most read ones.[1] And even today it can be said that her novels, as well as her name, i.e. pseudonym, are very well known to the Serbian public, owing to her novels that were screen adapted and very popular.

The expansion of the novel, as a literary genre, began in the early fifties of the twentieth century, and continues to this day. Dobrica Ćosic (1921–2014) is a writer who marked the post-war period. He wrote the novel *Koreni* (1954), which became the basis of the modern Serbian post-war novel and which led to great recognition. After the above mentioned as well as many other Serbian writers, the tradition of writing novels was continued by the new ones.

We could not find data on the historical development of Serbian novels that are classified as love novels, because such novels are difficult to find in the history of Serbian literature. Although there are many Serbian novels with the love motive as an important element, they are not defined as such. They are most often defined only as social, historical or novels about individuals, etc., or a combination of these genres. Today, love novels are quite popular and have large circulations, although most people consider them to be less valuable literary works. Mirjana Bobić Mojsilović (1959–), Jelena Bačić Alimpić (1969–) and Vesna Dedić (1967–) stand out among the most popular contemporary Serbian writers of love novels.[2]

The writers mentioned here, as well as others covered by this project, are known to the Serbian public (some to a wider and some to a narrow extent). What most of them have in common is that they received awards for their works and that their works have been translated into some of the foreign languages. As part of some, there are also awards named after them. These include: Borisav Stanković, Isidora Sekulić, Meša Selimović, Miloš Crnjanski, Momo Kapor, etc. Some of the awards the writers have received for the novels covered by this project are:

- Award of the Serbian Literary Cooperative for lifetime achievement — the novel *Došljaci* by Milutin Uskoković (1884–1915) (as the best manuscript in 1909);

1. Retrieved from Wikipedia (accessed on August 10, 2021.)
2. Retrieved from Wikipedia (accessed on August 10, 2021.)

- NIN's Award[3] which was first won by the novel *Koreni* by Dobrica Ćosić in 1954, *Roman o Londonu* by Miloš Crnjanski (1893–1977), 1971, *Sudbina i Komentari* by Radoslav Petković (1953–), 1993;
- Award for the most read book in Serbian libraries[4] — *Roman o Londonu* by Miloš Crnjanski in 1973, *Dorotej* by Dobrilo Nenadić (1940–), 1978.

When it comes to novels by female writers, there are other awards. The manuscript of the first novel by Grozdana Olujić (1934–2019) — *Izlet u nebo* won an award in 1957 at the competition of the Sarajevo National Education. Her second novel — *Glasam za ljubav* (1963) won the Zagreb Telegram Award for Best Short Novel. Marija Jovanović (1959–) was awarded the Women's Feather Pen for her novel *Spletkarenje sa sopstvenom dušom* (2000)[5]. Vida Ognjenović (1941–) received the award for the best book in the public libraries of Serbia for the novel *Preljubnici* (2006)[6]. And Jelena Bačić Alimpić was awarded the Golden Hit Libero for her novel *Ringišpil* (2010)[7].

4 About the multimedia project *The two of them*

The main topic of the project is the presentation of novels by Serbian writers whose characters are in romantic relationships. In addition, writers and novels were presented. These novels are diverse in type — not all of them are purely love novels, but they have a love motif as one of the main elements of the novel.

There are about 10 love novels,[8] along with novels listed in two genres, one of which is love, while there are more novels that belong to the social, historical and novels about individual people. Most of the novels from the 20th century can be found in the project, but there are also a few from the 19th and 21st centuries. There is a total of 25 novels, as many as participants in this project.

3. Prestigious Serbian, and formerly Yugoslav, literary award that is awarded every year for the best new Serbian (formerly Yugoslav) novel.

4. It was awarded by the National Library of Serbia from 1973–2004. years, with interruptions.

5. It was awarded by the women's magazine Politika Bazar, for the best novel or collection of stories by a female writer.

6. It was awarded by the National Library of Serbia from 2005–2013. years.

7. Recognition to the authors and publishers of the most sought-after books in the current year, which was awarded by the RTS Culture and Art Editorial Office and the Hit libris show.

8. Due to the lack of clear data on the genre, this number is not accurate.

The novels are presented in the form of web pages made by the final year students of the Department of Librarianship and Information Science of the Faculty of Philology in Belgrade, academic year 2019/2020. Assistance to students was provided by Prof. Dr Cvetana Krstev and Dr Branislava Šandrih Todorović, who were mentors on the project. This project is an integral part of the subject Multimedia Documents, which is part of the curriculum in the fourth year of undergraduate studies at the Department of Librarianship and Information Science.

5 Novels and writers covered by this project

The novels that the students chose to present were: ***Aristid i Natalija*** (1801) by Atanasije Stojković, ***Čudan svet*** (1869) by Jakov Ignjatović, ***Seoska učiteljica*** (1898) by Svetolik Ranković, ***Došljaci*** (1910) by Milutin Uskoković, ***Nečista krv*** (1910) by Borisav Stanković, ***Nove*** (1912) by Jelena Dimitrijević, ***Čedomir Ilić*** (1914) by Milutin Uskoković, ***Đakon Bogorodičine crkve*** (1919) by Isidora Sekulić, ***Plava gospođa*** (1924) by Milica Janković, ***Pokošeno polje*** (1934) by Branimir Ćosić, ***Srpska trilogija*** (1937) by Stevan Jakovljević, ***Pesma*** (1952) by Oskar Davičo, ***Koreni*** (1954) by Dobrica Ćosić, ***Izlet u nebo*** (1958) and ***Glasam za ljubav*** (1963) by Grozdana Olujić, ***Tvrđava*** (1970) by Meša Selimović, ***Roman o Londonu*** (1971) by Miloš Crnjanski, ***Dorotej*** (1977) by Dobrilo Nenadić, ***Una*** (1981) by Momo Kapor, ***Cvat lipe na Balkanu*** (1991) by Gordana Kuić, ***Sudbina i komentari*** (1993) by Radoslav Petković, ***Spletkarenje sa sopstvenom dušom*** (2000) by Marija Jovanović, ***Preljubnici*** (2006) by Vida Ognjenović, ***Ringišpil*** (2010) by Jelena Bačić Alimpić and ***Nikad nisam*** (2019) by Vesna Dedić.

Novels known to most students (before starting to work on this project) were: *Nečista krv*, *Koreni*, *Tvrđava*, *Cvat lipe na Balkanu* and *Došljaci*. Novels not known to most students were: *Spletkarenje sa sopstvenom dušom*, *Preljubnici*, *Aristid i Natalija*, *Nove*, *Seoska učiteljica*, *Plava gospođa*, *Čudan svet*, *Dorotej*, *Sudbina i komentari*. The writers known to most students were: Borisav Stanković, Isidora Sekulić, Oskar Davičo, Dobrica Ćosić, Meša Selimović, Miloš Crnjanski, Momo Kapor, Jelena Bačić Alimpić and Vesna Dedić. The writers not known to most students were: Marija Jovanović, Milica Janković, Dobrilo Nenadić, Atanasije Stojković, Svetolik Ranković and Jelena Dimitrijević.

5.1 Division of chosen novels by genre

As for the typological, i.e. genre divisions of the mentioned novels, it should be emphasized that such data are difficult to find. Novels such as *Izlet u nebo* and *Roman o Londonu* are not divided by genre, while the novel Dorotej is considered to be on the border of genres — it contains historical, love and psychological motives.[9] Of all the selected novels, the novels that are classified as love novels are: *Plava gospođa*, *Glasam za ljubav*, *Una*, *Spletkarenje sa sopstvenom dušom*, *Preljubnici*, *Ringišpil* and *Nikad nisam*. In addition to these, there are novels: *Došljaci*, *Čedomir Ilić* and *Đakon Bogorodičine crkve*, which also belong to other genres — both Uskoković's novels also belong to social novel, and Isidora Sekulić's novel also belongs to a religious novel. Other novels are classified into different genres (mainly social, historical, psychological and novels about individuals).

5.2 Dramatic adaptations of novels and/or novels that are screened

One of the important elements of the project is the information about the adaptations of the novels. As for the dramatization and screen adaptation of these novels, seven of them (out of 25) do not have their own theatrical or screen adaptation.[10] These include the novels: *Aristid i Natalija*, *Nove*, *Plava gospođa*, *Spletkarenje sa sopstvenom dušom*, *Preljubnici*, *Ringišpil* and *Nikad nisam*.

Fourteen novels have been theatrically adapted. These are: *Čudan svet, Seoska učiteljica*, *Došljaci*, *Nečista krv*, *Čedomir Ilić*, *Đakon Bogorodičine crkve*, *Pokošeno polje*, *Srpska trilogija*, *Pesma*, *Koreni*, *Izlet u nebo*, *Tvrđava*, *Roman o Londonu* and *Sudbina i komentari*. Among these plays, there are those with a title that differs from the original title of the novel. This includes performances, i.e. novels: *Vrela krv* (*Nečista krv*), *Sile* (*Pokošeno polje*), *Hromi ideali* (*Čedomir Ilić*), *Đakon* (*Đakon Bogorodičine crkve*; based on the motives of this novel and the short story *Ambicije, dim* by Isidora Sekulic) and *Na leđima ježa* (*Srpska trilogija*).[11]

Movies are made based on eight novels: *Pesma* (1961), *Izlet u nebo* (1962, known as *Čudna devojka*), *Glasam za ljubav* (1965), *Došljaci* (1969),

9. Data taken from the website of the online bookstore Laguna (accessed on August 15, 2021)

10. No information was found through an internet search.

11. Data taken from the website of the Museum of Theater Arts of Serbia (accessed on August 10, 2021)

Pokošeno polje (1980, known as *Beogradska razglednica 1920*), *Dorotej* (1981), *Una* (1984) and *Nečista krv* (1996 and 2021). The screen adapted novels are: *Čedomir Ilić* (1971), *Pesma* (1975), *Roman o Londonu* (1988), *Cvat lipe na Balkanu* (2011–2012) and *Koreni* (2018).[12]

6 Student assignments on the project

Before the creation of the multimedia document itself, it was necessary to accomplish several tasks. The first task for each student was to find books that would fit the topic of the project. All students, along with the professors, gave book suggestions while in the following class each student chose a book to read at home and make the MS PowerPoint presentation about it. The presentations contained basic information about the writers and books — the name and surname of the writer, the title of the novel, the year of its first edition, the biography of the writer, the list of characters and the description of the theme of the novel. Some of them also contained some interesting facts from the novels or information on whether those novels were screened or had theatrical adaptations. The next main task was to create web pages based on that, using HTML[13] language at home, along with CSS[14] for visual design and embellishment. Everyone designed and created a web page for the chosen novel independently (according to their wishes and knowledge).

In addition to these tasks, group tasks were also assigned — creating web pages for the entire project (designing and creating HTML code) and choosing music, i.e. a song that will bring the project to life. Two groups of students were formed for these group tasks. The group in charge of music consisted of students: Marija Despić, Ana Ivanović, Anica Jovanović, Sofija Hebar and Nikola Šibalić. The second group, which was in charge of coming up with the design and the technical production, consisted of students: Katarina Glavonjić, Magdalena Lukić, Jovan Janković and Stefan Joksimović. The group in charge of music first made a list of songs from YouTube, followed by voting for the song to be used on the site. An adaptation was chosen of

12. Screening data is mostly retrived from Wikipedia and the IMDb.

13. HTML (Hyper Text Markup Language) — A hypertext markup language, i.e. a descriptive language specifically designed to describe web pages. Retrieved from Wikipedia (HTML) (accessed on May 27, 2021.)

14. CSS (Cascading Style Sheets) — A cascading style sheet, a formatting language that defines the appearance of web page elements. Retrieved from Wikipedia (accessed on May 27, 2021.)

Richard Clayderman's composition *Love Story* from the film with the same title.

For better organization of tasks, in addition to email correspondence, Google Drive was also used. All the necessary documents and information about the assignments were placed there and were available to every student at all times. There was a table created (in the free online program Google Sheets) for the list of novels and student assignments, which was upgraded over time. There were, also, folders made for students to add their works to (presentations, HTML documents and accompanying elements that were used to create web pages, such as photos, etc.).

6.1 Data collection

Finding novels that are suitable for the topic of the project (those whose content encompasses the story of people in a romantic relationship) was not a very difficult task. At the beginning, there were small doubts about the topic itself, that is criteria for selecting novels and writers, because students were not quite familiar with such novels. But soon enough, with the help of the professor, a list of suitable novels was made. The proposals were discussed in the classes, and that is how the initial list was made, which was later upgraded by the students. Everyone had the task to find at least one novel and to upgrade the table with their suggestions (by adding the title, the name of the author and the year of the first edition of the novel). Thus the final list of novels was formed.

The reasons for choosing these novels were different — someone had heard of a novel that matched the topic or had already read it, someone asked for recommendations (mostly from family members), and some students found chosen novels on their own, by searching the Internet, while the others chose among the listed ones. Undoubtedly, the significance and popularity of most writers and novels helped students decide which work of Serbian literature they want to present.

After the final selection of novels, students read them and collected data on writers and novels afterwards. One of the main sources was certainly the chosen novels, which the students analyzed and presented (focusing on the topic of this project). In addition to the content of the novels, data from the books' covers were also used, as well as data from the prefaces and afterwords. Other sources of information were mainly Wikipedia, writers' websites and electronic newspapers (with various articles about writers and novels, which also contain information about plays, films, etc. as well as interviews with

writers) but also various other websites (mostly for finding photos, videos and other details and interesting facts).

7 Project websites

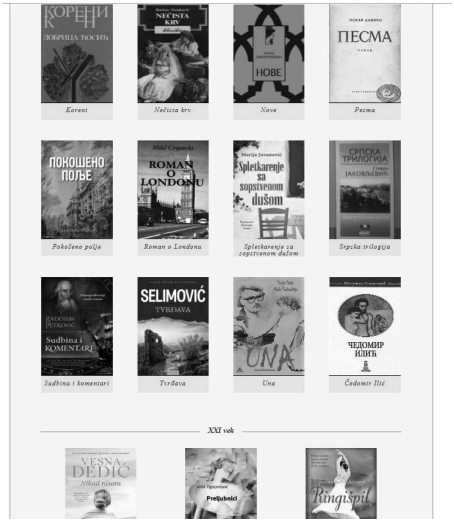


Figure 1. View of the first page where all the novels are listed

Multimedia project *The two of them* consists of three basic web pages. The first page of the project contains all the novels grouped according to the period of creation, from the 19th to the 21st century (Figure 1). There are photographs of books' covers used to show the novels, and below each there is its title. By clicking these photos, the pages created by the students open (figures2 and 3 that represent parts of HTML and CSS codes of one web page). Each of these pages is different in appearance and content from the others. The second page contains photos of all students who participated in the project, while the third page contains information about the project itself (Figure 4) — what is the purpose of the project, who are the authors of the project, what the web pages contain, while the names of the students who worked on creating the site (designing and creating the main pages) are placed in the bottom. These pages are enhanced by links to the website

of the Faculty of Philology in Belgrade and the website of the Department of Librarianship and Information Science, as well as links to the websites of Prof. Dr Cvetana Krstev and Dr Branislava Šandrih Todorović. An integral part of the site is an audio player with an adaptation of the composition *Love Story* by Richard Clayderman that the technical team (Jovan Janković and Stefan Joksimović) added to the site.

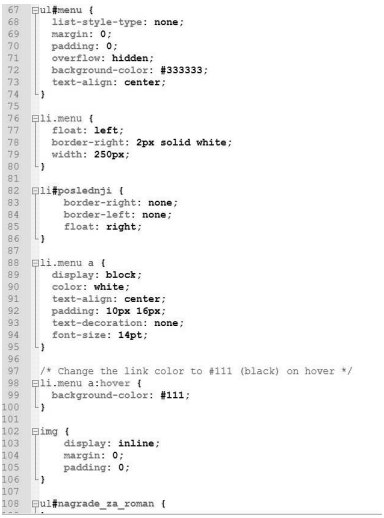
```
<body>

<div id="grid_container">

    <div id="header">
        <table>
            <tr>
                <td class="tabela_foto">
                    <img src="img/sudbina-i-komentari-korice-666x1024.jpg" alt="Naslovna korica" width="auto" height="390">
                    <p>Извор: <a href="https://www.originalmagazin.com/wp-content/uploads/2020/04/sudbina-i-komentari-korice-666x10
                </td>
                <td id="info">
                    <h1>Радослав Петковић</h1>
                    <h2>Судбина и коментари</h2>
                    <hr/>
                    <p id="naslovni_dodatak">Роман награђен НИН-овом наградом за књигу године</p>
                    <hr/>
                    <p>Приредио: Митар Зечевић</p>
                </td>
                <td class="tabela_foto">
                    <img src="img/1192255_Radoslav_Petkovic_08.jpg" alt="Portret autora" width="auto" height="390">
                    <p>Извор: <a href="https://www.vreme.com/g/images/1192255_Radoslav_Petkovic_08.jpg" class="naslovni_link">Vreme
                </td>
            </tr>
        </table>
    </div>

    <div id="nav_bar">
        <ul id="menu">
          <li class="menu"><a class="active" href="sudbina_i_komentari.html">Насловна</a></li>
          <li class="menu"><a href="oautoru.html">О аутору</a></li>
          <li class="menu"><a href="zanimljivosti.html">Занимљивости</a></li>
          <li class="menu" id="poslednji"><a href="https://sr.wikipedia.org/sr-ec/%D0%A0%D0%B0%D0%B4%D0%BE%D1%81%D0%BB%D0%B0%D0%B2_
        </ul>
    </div>

    <!-- Sadrzaj -->
    <br/>
    <div id="content">
        <p style="font-size: 18pt;">Историјски роман са елементима фикције.</p>
        <table>
            <tr>
                <td><p>Награде за роман:</p></td>
            </tr>
            <tr>
```

Figure 2. Display of HTML code of one page for one novel

The creation of (individual and main) web pages was not a difficult task for most of them. Students had little difficulty while working because they needed to revise HTML, and some had to learn more about it on their own (from lecture lessons), because their idea of web page design required higher knowledge. Therefore, it took a little more time and effort to make the web pages, but in the end, all students successfully completed their assignments within the set deadline (despite the outbreak of the Covid-19 pandemic).

8 Projects with similar content

Multimedia documents are an integral part of the curriculum from the school year 2009/2010. If we look back, we will notice that there is a connection between this year's and some earlier multimedia documents.



```
ul#menu {
    list-style-type: none;
    margin: 0;
    padding: 0;
    overflow: hidden;
    background-color: #333333;
    text-align: center;
}

li.menu {
    float: left;
    border-right: 2px solid white;
    width: 250px;
}

li#poslednji {
      border-right: none;
      border-left: none;
      float: right;
}

li.menu a {
    display: block;
    color: white;
    text-align: center;
    padding: 10px 16px;
    text-decoration: none;
    font-size: 14pt;
}

 /* Change the link color to #111 (black) on hover */
li.menu a:hover {
    background-color: #111;
}

img {
      display: inline;
      margin: 0;
      padding: 0;
}

ul#nagrade_za_roman {
```

Figure 3. Display of CSS code for one HTML page of one novel

Since the project *The Two of them* focuses on the novels of Serbian writers, the multimedia project of previous generations of students that is most similar to this one is the project *Literature on Film* (Меглић 2019), as well as the project *Pop Ćira i Pop Spira* (Коврлија, Тасић, and Топаловић 2012). These works include novels by Serbian writers, namely those that have been screened, and the screen adaptation of the novel is one of the additional elements in the project *The Two of them*. In addition to them, the project *Around the World in 80 Days* (Перић, Гогић, and Николић 2014) is similar by the topic because it also includes a novel that was screened.

9 Conclusion

Before starting to work on the project *The Two of them*, the students were familiar with many writers and novels that were on their list of novels suitable for this project. All of them easily got the necessary data (information about the writer, novel, photos, etc.), but minor problems arose while creating web pages because students had to revise HTML, which they studied in the second year of studies. A slight problem for some was the time needed to create the website, but in the end everything was completed without

Figure 4. Third page

major problems and within the scheduled deadline (although the Covid-19 pandemic somewhat changed the original work program plan).

This project was not too demanding for the students, they had the opportunity to use and improve their knowledge acquired at the university, but also to be creative in their own way. They discovered and read new books, learned more about writers, fictional characters or real people, as well as the society, place, period in which they lived, etc. Afterwards they presented it in front of the class, and then, after a while, they made websites for those novels. Finally, they united them all into one joint project, available to everyone on the Internet.[15]

The project combined research work (finding information on the Internet), knowledge of HTML, CSS and MS PowerPoint. All of this students could learn or improve during their studies at the Department of Librarianship and Information Science. PowerPoint lecture is part of the course Practicum of Informatics 1, which is held in the first year of undergraduate studies at the Department, while HTML and CSS are part of the course Digital Text 2, which is held in the second year. As far as research work is concerned, it has improved throughout academic studies through various tasks in many subjects, while it has reached even higher level within this subject.

15. On the website of Dr. Branislava Šandrih Todorović and the home page of Prof. Dr. Cvetana Krstev

Acknowledgment

We would like to express our special gratitude to professors Dr Cvetana Krstev and Dr Branislava Šandrih Todorović, who were always available for cooperation and assistance and who successfully guided us through this project in the subject Multimedia Documents, academic year 2019/2020.

References

Solar, Milivoj. 2005. *Teorija književnosti.* Zagreb : Školska knjiga.

Ђорђевић, Часлав, and Предраг Лучић. 2008. *Књижевност и српски језик : приручник за ученике четвртог разреда гимназије и средњих стручних школа.* 73–74. Нови Сад: Д. Ђорђевић, Венцловић.

Живковић, Драгиша. 1992. *Теорија књижевности са теоријом писмености : стилистика, књижевни родови и врсте, књижевни периоди и књижевни правци, о композицији и стилу писаних састава : приручник за наставнике и ученике.* Београд : Завод за уџбенике и наставна средства.

Коврлија, Дарја, Валентина Тасић, and Сузана Топаловић. 2012. "Израда мултимедијални документа „Поп Ћира и поп Спира", уједињено знање на заједничком студентском пројекту." *Инфотека* 8 (2): 86–89. http://infoteka.bg.ac.rs/pdf/Srp/2012-2/INFOTHECA_XIII_2_December2012_86-89.pdf.

Меглић, Катарина. 2019. "Израда мултимедијалног документа „Књижевност на филму"." *Инфотека* 19 (1): 96–107. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/160.

Перић, Катарина, Кристина Гогић, and Ана Николић. 2014. "Израда мултимедијалног документа „Пут око света за 80 дана"." *Инфотека* 15 (2): 56–62. http://infoteka.bg.ac.rs/pdf/Srp/2014-2/Srp2014-2INFOTHECA_XV_2_april_56-62.pdf.

Тартаља, Иво. 1998. *Теорија књижевности : за средње школе.* Београд : Завод за уџбенике и наставна средства.

EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School

UDC 81: 37.018.53
DOI 10.18485/infotheca.2021.21.1.7

ABSTRACT: The first training school organized by the *NexusLinguarum* COST Action was held on February 8-12, 2021 and was aimed at students, academics, and practitioners wishing to learn the basics of Linguistic Data Science. During the training school, the participants were introduced to a wide range of topics: from Semantic Web, RDF and ontologies, to modeling and querying linguistic data with state-of-the-art ontology models and tools. The training school was organized under the umbrella of the EUROLAN series of summer schools and was hosted virtually (online) by several institutions: the Romanian Academy, the Research Institute for Artificial Intelligence in Bucharest and the Institute of Computer Science in Iași, as well as the "Alexandru Ioan Cuza" University of Iași, Romania. The training school was attended by 82 participants.

KEYWORDS: linguistic data science, linked data for linguistics, language data, NexusLinguarum, COST action, EUROLAN, training school.

PAPER SUBMITTED: 30 June 2021
PAPER ACCEPTED: 13 July 2021

Milan Dojchinovski
milan.dojchinovski@fit.cvut.cz
CTU in Prague
Prag, Czech Republic
InfAI at Leipzig University
Leipzig, Germany

Julia Bosque Gil
jbosque@unizar.es
Jorge Gracia
jogracia@unizar.es
Aragon Institute of Engineering Research (I3A)
University of Zaragoza
Zaragoza, Spain

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

1 Introduction

NexusLinguarum - European network for Web-centered linguistic data science, COST action CA18029[1] - was launched at the end of October 2019. The goal of the *NexusLinguarum* action is to promote the study of linguistic data science, for which the construction of an ecosystem of multilingual

1. NexusLinguarum

and semantically interoperable linguistic data is required. Training schools are one of the means for reaching this goal, and therefore the *NexusLinguarum* core team organized the Introduction to Linked Data for Linguistics online training school[2] that took place from February 8 to 12, 2021. The training school was aimed at promoting and teaching the basics of linguistic data science and the related technologies to people from the academia and the industry. It was organized under the umbrella of the EUROLAN series of summer schools, which was established in 1993 and covers topics that are particularly relevant to the fields of computational linguistics and natural language processing (NLP). The goal of this 15th EUROLAN School was to bring together scholars, teachers and students of linguistics, NLP and information technology to discuss the principles and best practices for representing, publishing and linking linguistic data and the issues that constitute the building blocks in the envisioned multilingual and interoperable web-oriented ecosystem. The present contribution summarises the organisation, content and results of this training school and is based on Deliverable D1.1[3] of the Action.

2 The training school program

The training school has been developed for newcomers as well as for those already having basic knowledge in the fields covered. The school provided a comprehensive introduction to the methodologies for representing linguistic resources using semantic web technologies, together with the means to extract knowledge from language resources and exploit it using semantic web query languages and reasoning capabilities. The topics addressed in the school were the following:

- Semantic Web and Linked Data[4] (Berners-Lee et al. 2006);
- Ontologies: RDF (Resource description framework), RDF Schema (Resource Description Framework Schema, variously abbreviated as RDFS, RDF(S), RDF-S, or RDF/S), Web Ontology Language (OWL),[5] etc.);
- SPARQL query language- a semantic query language for databases able to retrieve and manipulate data stored in the RDF format;

2. EUROLAN
3. Deliverable D1.1
4. Introducing Linked Data and the Semantic Web
5. OWL

- Metadata: DCAT (Data Catalog Vocabulary),[6] VOID (RDF Schema vocabulary for expressing metadata about RDF datasets, etc.);
- RDF transformation and validation; (Cimiano et al. 2020)
- Linguistic linked data; (Chiarcos et al. 2013)
- Lemon-OntoLex[7] (McCrae et al. 2017; Declerck, Tiberius, and Wandl-Vogt 2017; Stanković et al. 2018)
- Linguistic linked data generation; (Cimiano et al. 2020)
- Corpora and linked data; (Chiarcos 2012)
- Linguistic annotations; (Fäth et al. 2020)
- NLP Interchange Format; (Hellmann et al. 2013)
- Tools and applications of linguistic linked data. (Declerck et al. 2020)

The first day started with an opening session and a brief introduction to Linguistic Linked Data (LLD), followed by an introduction to Linked Data and RDF dedicated sessions. The second day covered topics related to ontologies, including modelling knowledge with ontologies, OWL and SKOS[8] knowledge representation languages, reasoning of knowledge, and a hands-on session using the Protégé[9] ontology editor. The third day was dedicated to the topics related to representing and querying lexical data with dedicated sessions on the OntoLex-Lemon model and the SPARQL querying language. The fourth day included sessions which gave an overview of other linguistic and metadata vocabularies and the VocBench platform[10] (Stellato et al. 2020) modelling linguistic datasets. In the afternoon, an online social event was organized where the participants could remotely see the beauty of the Romanian culture, traditions and nature. The fifth day comprised three parallel sessions on different topics:

(i.) LLD Generation/Transformation and Linking,
(ii.) Annotations (NIF, Web Annotation) (Hellmann et al. 2013), and
(iii.) OntoLex extensions: *vartrans* for representing translations and term variants (based on the *lemon* translation module, (Gracia et al. 2014)), *lexicog*[11] – lexicography module (Bosque-Gil, Gracia, and Montiel-

6. Data Catalog Vocabulary (DCAT) - Version 2
7. Lemon - Lexicon Model for Ontologies; Lexicon Model for Ontologies: Community Report, 10 May 2016
8. SKOS Simple Knowledge Organization System - home page
9. Protégé
10. VocBench: A Collaborative Management System for OWL ontologies, SKOS(/XL) thesauri, OntoLex-lemon lexicons and generic RDF datasets
11. The OntoLex Lemon Lexicography Module

Ponsoda 2017), *FrAC*[12] – frequency, attestation and corpus Information (Chiarcos et al. 2020).

Finally, the training school ended with a closing session where an ontology of participants, lecturers and organizers was presented, illustrating many of the representation mechanisms explained throughout the week.

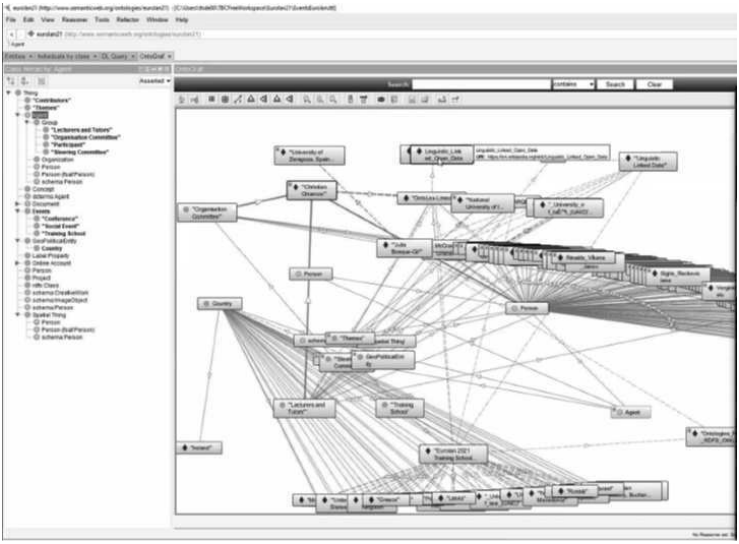


Figure 1. Ontology of the training school.

.

Each of the organized sessions was accompanied by a hands-on session and an exercise session. During the hands-on session, the lecturers proposed an exercise and offered a step-by-step walk-through for the participants to understand the methodology leading towards the solution. They also introduced the basic technology needed. Then, during the exercise session, the participants were asked to work on a particular task like the cases presented during the hands-on session, thereby becoming familiar with the technology introduced in a practical setting. As these sessions were graded in terms of complexity, starting with the basic notions, and building on to present more specific topics in a detailed fashion on the last day, the participants had

12. FrAC – Frequency, Attestation and Corpus Information - Ontology-Lexica Community Group

a chance to acquire a solid foundation before moving onto more complex sessions. The official program of the school is available online.[13]

As a follow up, the JeRTeh[14] Language Resources and Technologies Society set up a local installation of VocBench[15] and, apart from JeRTeh members, it was used by students and teachers of the Intelligent Systems PhD program[16] at the University of Belgradefor the subjects Knowledge representation and Semantic web. The Lemon-OntoLex Frac module was used for representation of the entries from the lexicon used for abusive speech detection with attestations from the Twitter corpus with annotation of abusive spans (Jokić et al. 2021).

3 Organization

Due to the COVID-19 pandemic and current travel restrictions in Europe and beyond, the training school was held online. Following on the almost three decades long tradition of EUROLAN, which is known for academic program excellence and camaraderie among professors and students, a range of virtual activities were carried out in addition to holding online classes, with the aim of providing cultural experiences and discoveries, in addition to closer interaction. Attendance was online and free of charge, requiring pre-registration.

All the sessions were hosted using a videoconferencing platform. For the hands-on sessions, several breakout (virtual) rooms were made available where the participants could work on the assignment in smaller groups. To encourage participants to ask questions and get in touch with each other, the organizers set up a Slack[17] channel, as a collaboration hub where lecturers and participants could clarify any doubts. The total number of participants was 82, 52 female and 30 male, including 4 participants from Serbia.

Various types of materials were generated for the training school, including presentations (slides)[18] and exercises[19] accompanied by code and data

13. School Program
14. Jerteh
15. VocBench installation
16. Intelligent Systems PhD Program
17. Slack, The School channel
18. Presentations
19. Exercises

examples.[20] All the materials were published online and made available for free.

4 Summary

The training school provided valuable knowledge and trained many computer scientists and linguists on how to work with and benefit from linguistic linked data. This was the first training school organized by the *NexusLinguarum* COST Action as one of a series of training events that are planned to take place. It aimed to serve as an introduction to the topic of linguistic data science and build the basis for the audience necessary for attending future training schools on more advanced topics for the duration of the COST Action. All the materials created during the training school are publicly available and can be further used by the community. During the closing session, the organizers provided participants with a survey form to gather feedback on both organizational and academic aspects of the school. The results have shown that the disciplines of the humanities/linguistics/lexicography had a higher representation among participants than computer science, and that the school was well-focused, well-balanced topic-wise and well organized. Theory sessions, tutoring, and the opportunities to learn were very highly evaluated. On the other hand, due to the virtual mode, there is still room for improvement in practical sessions, social event organization and opportunities to network. The knowledge and skills acquired there will improve the development of Serbian linguistic resources and help to publish more resources as linguistic linked data.

Acknowledgment

This paper is supported by the COST Action CA18209 - *NexusLinguarum* "European Network for Web-centred Linguistic Data Science".

References

Berners-Lee, Tim, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. 2006. "Tabulator: Exploring and analyzing linked data on the semantic web." In *Proceedings of the 3rd international semantic web user interaction workshop,* 2006:159. Athens, Georgia.

20. Supporting material

Bosque-Gil, Julia, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. "Towards a Module for Lexicography in OntoLex." In *LDK Workshops,* 74–84.

Chiarcos, Christian. 2012. "Interoperability of corpora and annotations." In *Linked Data in Linguistics,* 161–179. Springer.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Chiarcos, Christian, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. "Towards open data for linguistics: Linguistic linked data." In *New Trends of Research in Ontologies and Lexical Resources,* 7–25. Springer.

Cimiano, Philipp, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. "Converting language resources into linked data." In *Linguistic Linked Data,* 163–180. Springer.

Declerck, Thierry, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, et al. 2020. "Recent developments for the linguistic linked open data infrastructure." In *Proceedings of the 12th LREC,* 5660–5667.

Declerck, Thierry, Carole Tiberius, and Eveline Wandl-Vogt. 2017. "Encoding lexicographic data in lemon: Lessons learned." In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets. CEURS,* vol. 8.

Fäth, Christian, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 2020. "Fintan-flexible, integrated transformation and annotation engineering." In *Proceedings of the 12th LREC,* 7212–7221.

Gracia, Jorge, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-De-Cea. 2014. "Enabling Language Resources to Expose Translations as Linked Data on the Web." In *Proceedings of the 9th LREC,* edited by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-8-4.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. "Integrating NLP using linked data." In *International Semantic Web Conference,* 98–113. Springer.

Jokić, Danka, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih. 2021. "A Twitter Corpus and lexicon for abusive speech detection in Serbian." In *Proceedings of the 2021 Language, Data and Knowledge (LDK), 1-3 September in Zaragoza, Spain.*

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stellato, Armando, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem Van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costetchi, et al. 2020. "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons." *Semantic Web* 11 (5): 855–881.

COBISS Meet 2020 – Online Conference Presentation of New Approaches for All participants in the COBISS.net Network

UDC

Ana Vasiljević
ana.ivkovic.va@googlemail.com

Jelica Zeljić
jelicazelj1971@gmail.com

National library
Osečina, Serbia

PAPER SUBMITTED: 18 March 202
PAPER ACCEPTED: 1 July 202

The Institute of Information Sciences in Maribor (IZUM) organized COBISS Meet 2020 conference from 1st to 3rd of December 2020. The gathering was organized online because of the Covid-19 pandemic. The conference took place live on the internet platform https://hopin.to/events/cobiss-meet-2020. The virtual scene was a great opportunity to discuss ideas and joint plans for the future, contact with lecturers, colleagues and other participants of the event. The event was attended by over 600 participants from more than 10 countries. The central topic was the presentation of new approaches and innovations that have recently been implemented by IZUM for all participants in the COBISS.net network.

Welcome speech, lectures and a discussion with lecturers from Slovenia were scheduled for the first day of the conference. In his welcome address, Mr. Alen Bošnjak, PhD, Principle of IZUM, referred to the ongoing circumstances in the world that, however, did not prevent us all to continuously fulfill our mission. Even in such circumstances, we must enable our customers our full services. Afterwards, other lectures were announced and it was pointed out that everyone would have the opportunity to exchange views with lecturers and other participants after each session, which was an excellent chance to exchange ideas and make joint plans for the future.

In his welcome speech, Mr. Jure Gašparič, PhD, State Secretary in Ministry of Education, Science and Sport, pointed out the importance of IZUM and COBISS throughout his educational and research career in the field of librarianship, whose activity is not only recognized in Slovenia but also

throughout the region, as well as by UNESCO. COBISS and SICRIS are two very important parts in the mosaic of the Slovenian Development Strategy 2030, which is focused on the quality of life of all because the educational strategy that connects science and economics is of key importance for the transfer of knowledge. He also commented that the COBISS Meet 2020 event was probably the largest project in the field of science that IZUM organized in recent years in cooperation with Slovenian universities. The Academic Digital Collection of Slovenia is a relevant national information resource for researchers. It will be a unique platform for accessing full text versions of electronic information sources to which universities and their libraries are subscribed individually, which will significantly improve the overall user experience. IZUM is preparing a common open science national repository, dubbed dCOBISS. It enables the storage of open access researchers' publications and it provides project details. The upgrade of the national infrastructure – HPC RIVR – supercomputer will start working in the spring 2021 and it will strengthen the national high performance computing capacities needed for innovation in both research and economics.

In the presentation that followed, Mr. Bahan Neupane, the representative of UNESCO, talked about the cooperation between UNESCO and IZUM. The UNESCO Executive Board began considering the work of IZUM eight years ago. The agreement between the government of Slovenia and UNESCO was signed in September 2012 for a period of six years, however the center became functional on January 14$^{\text{th}}$ 2020, under the auspices of UNESCO. This effort will have a direct impact on the exchange of knowledge between different countries. Mr. Bahan Neupane also added that IZUM should continue to work in the key area of influencing the policies of how knowlegde gets produced, processed and archived, as it must bridge in between where the knowledge exists and where the knowledge lacks. He stressed that IZUM should disseminate through workshops, various types of brainstorming and knowledge exchange processes. The things done by IZUM are example of good practice for countries in the region and beyond. He specially emphasized and praised the preservation of multilingualism within the COBISS.net system.

In the presentation of Mr. Miran Petek we had the opportunity to hear about the digital repository COBISS, or dCOBISS for short, which is a new product of IZUM, installed and available to all libraries in the COBISS system. It is the latest IZUM web application intended for storing digital content used in libraries. Afterwards, Mr. Petek explained the connection with the Slovenian part of the repository which is connected to Slovenian

libraries. The application for this web application is autonomous and closely integrated with COBISS3-cataloging, COBISS+ and Sicris. All libraries in COBISS.net will have access to the central repository; librarians from different systems will be able to search the repository while librarians from a local system of a member state will be able to store and search records and files in their central repository. First step in the process of work of librarian is creation of bibliographic record in COBISS3/cataloging while copyright data and open access digital content files are added to bibliographic record in the dCOBISS. Slovenian research agencies as well as other research academic institutions support the preparation for the analysis of open access to a local repository, which has yet to become functional. The data from a local repository are aimed to be entered into the academic repository, which will be part of dCOBISS and in open access. Also, all libraries will be able to access the dCOBISS test environment where librarians with username and password may practice, test and learn and try out new upgraded features.

Ms. Janita Tacer Slana presented new information about COBISS+, such as ordering of articles stored in COBISS+ in the last two years. The development, upgrade and update of COBISS with new functions is constant and always taking into account the remarks of librarians and users. Ms. Tacer Slana presented the new characteristics of the My COBISS profile, concerning the production of the lists of search results and the limitation of search aspects. The last significant upgrade which refers to the adaptation for the blind and visually impaired, as well as the user interface for researchers were also presented. All portals work on the common COBISS+ software platform. One of advantages for users is that when they enter a search term, they search both COBISS and e-resources in full text. To improve customer service the Excel export option is included. Ms. Janita Tacer Slana also presented ways to reach different user groups in faster, easier and more comprehensive manner.

In the presentation of Ms. Tadeja Brešar, Head of Bibliographic Control at IZUM, we had the opportunity to get familiar with the general list of subject headings of COBISS. She indicated that there are 5.5 million records in the joint catalog and approximately 83% of these records contain at least one subject heading. Half of the records were created in the National and University Library (NUK) and other Slovenian libraries, and only 6% of the records contain a subject headings from another lists of subject headings (Library of Congress Subject Heading, REMEAU, etc.). The list of subject headings was created in the form of authority database that enables a unique system of naming subject headings for the entire COBISS system. This means that all

libraries could use the same list of these subject headings. The database enables the controlled entry in the field of the subject entry, which means that incorrect subject headings cannot appear nor can different catalogers express them differently. Authority control allows control of synonyms, which means that all authority records contain both normative and variant forms; besides it is required for all catalogues to use only the normative form of the subject entry when adding it. A general list of subject entries is structured and the entries are hierarchically related, both with superordinate and subordinate terms.

The second day of the conference started with the presentation of Ms. Tanja Turšek who works in the department for development of local applications COBISS3. Her presentation showed some new functions and updates that are built into COBISS3. She presented some functions of the most frequently used Loan software module, like updates related to the faster access to most often used material, member search by IDs or nemas, access to textbooks, e-book lending, booking via COBISS+ or mCOBISS.

Ms. Tanja Žuran Putora presented the new generation of inventory, the online inventory, that was implemented in 2019; it is performed in the COBISS3 production environment. The development took place in two steps, the solution for the revision of monographs was prepared at first, after which the possibility for the revision of serial publications was added. The application enables finding material on the shelf based on inventory numbers by using bar code readers.

In the mCOBISS application development, Mr. Boštjan Batič has the role of a programmer but he also works on the promotion and training of users for the use of the mobile application. This presentation referred to the unique mCOBISS application within the COBISS.net network. In 2013, IZUM started developing the mCOBISS mobile application, the project was co-financed by the European Union under the auspices of the Ministry of Education, Science and Sport of Slovenia. The aim of submitting the project was to encourage the development of a wider selection of e-services and mobile applications in various forms of information society, as well as to establish new technological concepts that will contribute to the transition to a modern and efficient knowledge society.

Mr. Sergey Lach presented COBISS as a social network, as he works in IZUM on public relations, training and building user relations, and is a member of editorial boards of several IZUM social media. Lach pointed out that the Agreement on the COBISS.net network and the exchange of bibliographic records between the national library information systems in the

countries of the Western Balkans was signed in 2003. COBISS.net develops library information system, connects eight countries and a total of 1400 libraries; over 5800 librarians are connected and cooperate. The network uses information technologies in 8 national languages and 2 scripts, which is a huge network of librarians and library users who can connect to social networks run and maintained by IZUM.

On the third day, Mr. Miran Petek presented the ADZ project related to the Academic Digital Collection of Slovenia. This project started last year when the first version of the so called *discovery* tools on the COBISS+ platform was installed. COBISS+ was first presented at the COBISS conference in 2014, while it was installed in 2016, and the previous system COBISS/OPAC was replaced by the COBISS+ platform. The basic platform that uses the functionality of the *discovery* tool for the research and academic environment was thus set up. This portal is intended to become the portal of the Academic Digital Collection of Slovenia while the development of a special part of COBISS+ intended for the research is planned.

Ms. Romana Muhvič Šumandl, head of the Service Management Sector, presented the new customer support functions. The project of including school libraries in the COBISS system was completed in 2018 and presented to the Associations of Library Societies. For the needs of school libraries certain conversions had to be done as some school libraries worked using different systems, such as WinKnj or School Library. The software upgrade with the option Material in quarantine, which is related to the work of libraries in the situation caused by the covid-19 virus pandemic was presented afterwards. Ms. Muhvič Šumandl mentioned that the enrollment of users via the Internet increased due to the pandemic. She also added that the commercial Oracle database was replaced by the relational open source database PostgreSQL. An interesting fact is that a total of 95 libraries joined COBISS.net in two years.

Mr. Bojan Štok, the Head of Software Development, introduced us to the architecture of the COBISS system and presented the transition to an open source platform. He presented the architecture of the COBISS system as it is today, the technology changes involved in the development of COBISS2 and COBISS3, as well as the future technology that will be used in the development of COBISS4 software. The container technology will be used instead of virtual servers. Technological changes will depend on the development of technology in the future. JakartaEE and MicroProfile technology, which are the successors of JAVE EE enable faster application development and are planned to be used.

Ms. Gordana Mazić presented concept and preliminary results of the analysis performed within the task of creating normative records for works within the COBISS system. She pointed out that when creating an authority record for a certain title, i.e. work, it is necessary to link it with records for all its editions and translations into different languages. An authority record for a work is related to authority records for authors as well as to the bibliographic records for its various editions and translations. Authority records for works are like a jigsaw puzzle that builds a broader picture of the works itselves. Keeping such normative records puts intellectual and artistic content at the highest level of abstraction.

The lectures that were held every day, lasted in total two hours and were followed by discussion. For better understanding all lectures were translated into three languages. The moderators were: Mr. Boštjan Batic and Mr. Branko Kurnjak.

This virtual gathering was a great opportunity to discuss ideas and joint plans for the future. In addition to monitoring professional work, exchange of opinions, examples of good practice and achievements with lecturers, colleagues and other participants, this event was also a great opportunity for social networking and sharing of more than 1500 comments, greetings, questions and answers lined up on the conference page during the event. The presentation of new approaches and innovations, which IZUM introduced in the last period, were well received, which was reflected in the discussions with the lecturers. The interest in participating in this event was great. The welcome speech and all lectures, including discussions are available on the You Tube channel COBISS - the path to knowledge, IZUM Maribor.[1]

1. Conference video material

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).