Impressum

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

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

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

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

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

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

Belgrade, Vol. 18, No. 2, December 2018

WEB PORTAL:
Vukosav Sredojević
University Library “Svetozar Marković"

LECTORS FOR ENGLISH:
Jelena Jaćimović, PhD, *Faculty of Dental Medicine, University of Belgrade*
Jelena Mitrović, PhD, *Department of Informatics and Mathematics, University of Passau*

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

DESIGN AND PREPRESS:
Aleksandar Milošević, Branislava Šandrih
and **Info***theca* Team

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

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

JOURNAL REDACTION:
Journal Infotheca
11000 Belgrade, Bulevar kralja Aleksandra 71

+381 11 3370–211
infotheca@unilib.rs

PRINTED BY:
Mamigo plus
Belgrade

Journal is published twice a year

Contents

Scientific papers

Professional papers

Reviews

Aligned Parallel Corpus for the Domain of Management: Preparation and Potential Applications

UDC 81’322.2
UDC 81’373:005
DOI 10.18485/infotheca.2018.18.2.1

Jelena Anđelković
plecasj@fon.bg.ac.rs
University of Belgrade
Faculty of
Organizational Sciences

Danica Seničić
danica.senicic@gmail.com

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

University of Belgrade
Faculty of Mining and Geology

ABSTRACT: The paper presents an aligned parallel English-Serbian corpus for the domain of management. This is a larger text collection available through Biblisha, an aligned collection search tool. In addition to describing the content of the corpus, the selection criteria and the text compilation process, the paper also illustrates the possible applications of the corpus by using corpus analysis examples in several areas of linguistic studies, i.e. in terminological, comparative and contrastive studies of language for specific purposes (LSP), specialized and academic discourse, translation process, and in the teaching and learning of English for specific purposes (ESP).

KEYWORDS: parallelized corpus, domain corpus, terminology, terminology unit, management

PAPER SUBMITTED: 29 December 2017
PAPER ACCEPTED: 15 November 2018

1 Introduction

In older linguistic literature, a corpus was defined as a collection of texts chosen to represent a language, a dialect or a language subsystem, and used for the purpose of linguistic analysis (Pearson, 1998, 42). More recently, along with the development of computational linguistics and natural language processing tools and systems, the definition of corpora has been changing, too. Today, a corpus is mostly seen as a *machine-readable text collection* (McEnery and Wilson, 2001, 177), or *a linguistic collection of texts in electronic form, selected on the basis of external criteria to represent, in the best* *way possible, a language or language variety as a source of data for linguistic research* (Sinclair, 2005). McEnery, Xiao and Tono (McEnery et al., 2006, 13) believe that corpora are *machine-readable collections of authentic texts* (including speech transcripts) *sampled* in such a way as to be *representative* of a particular language or a language variety.

There are several typologies of language corpora; and the choice of a corpus type depends on the purpose of linguistic analysis that we intend to conduct. While a *general language corpus* is a collection of written and / or spoken language that represents (or should represent) a particular language as a whole, a *specialized corpus* (also known as a *corpus for specific purposes* or a *domain corpus*) is an electronically accessible collection of texts that represents a specialized area of communication, and is representative of a specific domain of language use. Specialized corpora are often composed of *genre-specific* or *domain-specific texts*, i.e. they are representative of only one specific scientific and professional domain or a discipline.

Parallel and aligned parallel domain corpora (i.e. corpora for specific purposes) are of particular importance for terminological, contrastive and comparative language studies. A *parallel corpus* is a bilingual or multilingual collection that contains equivalent texts (an original and its translations) in two or more languages (Tognini-Bonelli, 2001, 6)(Pearson, 1998, 47). In some cases, parallel corpora can contain texts in only one language, i.e. pairs or groups of different translations of the same text into a same language. In addition to being parallel (containing translation equivalents), aligned parallel corpora also is also aligned at a paragraph level, sentence level or the level of individual words.

Aligned parallel corpora for specific purposes are primary linguistic resources for multilingual processing in computational linguistics. These corpora enable systematic processing of large amounts of terminological information and automatic or semi-automatic extraction of terms and their equivalents in a foreign language. Lexical knowledge gained by using aligned parallel corpora is of particular importance in natural language processing (NLP), e.g. for the development of software systems and tools for machine translation, creation of bilingual electronic terminological glossaries, vocabularies, lexicons and terminology bases data, etc. (Véronis, 2013, 238). Aligned parallel text collections that include the Serbian language are relatively infrequent and not easily available, mostly due to the fact that adequate texts (original and translation equivalents) are often difficult to obtain.

First scientific papers regarding aligned parallel text collections that include Serbian language text equivalents appeared in early 1990s (Krstev C., 1994). In late 1990s, Plato's *Republic* was aligned in 17 languages (Vitas, 1998b), and Orwell's 1884 in seven languages (Vitas, 1998a), both including Serbian.

Evroteka[1] is a bilingual, English-Serbian, corpus of legal texts or excerpts created during the process of translating European Union legal texts into the Serbian language. The Communication Sector of the Serbian Ministry for European Integration has been in charge of its maintenance since 2009. The translation of the legal texts, as well as the creation of this corpus, is based on SDL Trados. [2] and its Translator's Workbench tool

As for the number of languages included, the multilingual aligned parallel corpus "MULTEXT-East "1984" annotated corpus 4.0". This corpus, available in CLARIN.SI repository, has been developed as a part of *the MULTEXT East - Multilingual Text Tools and Corpora project for Eastern and Central European languages*(Erjavec and Ide, 1998). The MULTEXT-East "1984" corpus consists of George Orwell's novel "1984" in English (original) and its translations into twelve Eastern and Central European languages (Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak and Slovenian). The texts have been aligned at sentence level, while lemmas and morphosyntactic descriptions have been manually validated (Erjavec et al., 2010). SrenWac[3] (Ljubešić et al., 2016), an aligned parallel corpus of Serbian and English consists of electronic texts taken from the .rs domain and generated automatically using the Spidextor. [4] tool

Steps and approaches to term extraction differ (Pazienza et al., 2005), for example, describes the following: 1) the use of statistical measures for selecting relevant terms from the list of term candidates, 2) identifying and recognizing terminological expressions using only linguistic approach and filtering specific syntactic terminological patterns, and 3) hybrid approaches that merge the previous two, taking into account syntax properties and statistical measures for term recognition (Siddiqi and Sharan, 2015) lists two more approaches for term extraction: 4) the use of machine learning methods, and 5) the use of domain-specific knowledge resources (e.g. ontologies). This paper uses a hybrid approach that combines syntactic pattern recognition and statistical measures, described in (Stanković et al., 2016a).

[1] Evroteka (on-line)
[2] SDL Trados (on-line)
[3] SrenWac (on-line)
[4] Spidextor (on-line)

Aligned and electronically accessible Serbian-English corpora are developed by the Language Technologies Group at the Faculty of Mathematics, University of Belgrade, and by the Society for Language Resources and Technologies (JeRTeh),[5] with most members from the University of Belgrade. One of the aligned corpora developed by these two groups is *SrpEngKor*, a Serbian-English aligned corpus, containing texts from different genres (e.g. literature, journalism, law, medicine, education, etc.) and that are segmented and aligned (in most cases) at sentence level.[6]

Jerteh has also developed *Biblisha*, an aligned collection search tool. A detailed description of its implementation and use can be found in (Stanković et al., 2016b). The *Biblisha* collection currently contains several text collections, with each collection covering one or several related domains, e.g. librarianship and informatics, mining and geology, dentistry, architecture and urban planning, etc.

The following chapters describe the process of compilation and processing of the management - domain text collection. This is one of the larger collections of aligned texts available through *Biblisha*. The paper also highlights the potential uses of this collection in various types of linguistic research. The emphasis is not placed on the primary purpose of aligned parallel domain corpora (in computer linguistics and terminology management), but on less explored uses, i.e. in applied linguistics, comparative and contrastive studies of terminology, language for specific purposes and specialized discourse.

2 Aligned parallel corpus for management domain

2.1 Corpus contents

The aligned parallel English-Serbian specialized corpus for the domain of management consists of scientific papers published in the international journal *Management: Journal for Theory and Practice of Management*. The corpus is therefore both domain-specific and genre-specific. The corpus consists of 17 journal issues published between 2008 and 2012, with the total of 181 research papers containing approximately 30,000 sentences and more than 600,000 words per language (more precisely, 611,651 words in the Serbian part of the corpus). A more detailed overview of the corpus contents is presented in the table 1:

[5] JeRTeh (on-line)

[6] SrpEngKor (on-line)

issue (number/year)	No. of papers (per language)	No. of sentences (Serbian)
47-48/2008	12	2.187
49-50/2008	14	2.097
51/2009	9	1.503
52/2009	9	1.575
53/2009	10	1.194
54/2010	10	1.817
55/2010	10	1.750
56/2010	10	1.648
57/2010	10	1.502
58/2011	10	1.475
59/2011	10	1.501
60/2011	11	1.426
61/2011	14	2.301
62/2012	12	2.297
63/2012	10	1.815
64/2012	10	1.655
65/2012	10	1.583
$\sum$	181	29.326

Table 1. Corpus contents

The international scientific journal *Management: Journal for Theory and Practice of Management* is published quarterly by the Faculty of Organizational Sciences, University of Belgrade, a leading academic institution for this field in Serbia. The magazine aims to "enable relevant information exchange and communication between scientists, researchers, managers, and people in different business areas, coming from universities, institutes, companies and public services". [7] In the time period covered by our corpus, it was listed as a journal of national importance (M51category) by the Ministry of Education, Science and Technological Development of the Republic of Serbia. All the papers accepted by the journal are available on the journal's website, both in English and in Serbian. [8]

[7] Management: Journal for Theory and Practice of Management (on-line)

[8] Management, archive (on-line). Even though the papers are publicly available, a permission for their use in our aligned parallel corpus was obtained from the for-

2.2 The corpus: advantages and limitations

Authorship The papers in the corpus are either submitted by a single author or ther are composite texts written by two or more authors. The authors are either researchers into the domain of management and academic community members, or representatives of national, regional, or international companies and institutions. The relevant metadata shows that out of the total 181 papers, 21 papers (11.6%) were submitted solely by foreign authors (outside the territory of the former Yugoslavia), while the remaining 160 papers (88.4%) were either authored by Serbian authors, authors from the region of former Yugoslavia, or in co-authorship between the two groups.

The available metadata and the information from the journal's website do not indicate which of the two languages (English or Serbian) the papers were originally written in, and whether the paper was translated by the authors themselves or by professional translators; the assumption is that papers by foreign authors were originally written in English, and then translated into Serbian, while the authors from Serbia and the region did the opposite. Although such a text composition can significantly affect the quality, precision and monosemy of contained terminology, we believe that it can provide a better picture of terminological and other types of linguistic variation conditioned by pragmatic or sociolinguistic factors. For this reason, the papers were not selected with regard to authorship, i.e. to the language they were originally written in.

Pragmatic factors for text selection The choice of texts for the aligned parallel management corpus was primarily determined by pragmatic factors: the electronic availability of adequate texts, and the intended purpose of the corpus: linguistic and terminological research in this and related scientific and professional domains. Both these factors have certain advantages and limitations.

Corpus size. The corpus size (approximately 600,000 words per language) resulted from the availability of translated management-related research papers in Serbian and English. Even though corpus linguists do not fully agree on the optimal size of a corpus (Roe, 1977; Fang, 1993; Gledhill, 2000), i.e. on the ideal size of a specialized corpus (Flowerdew, 2004, 18), we believe that the management corpus presented here is adequate for a significant number

mer editor-in-chief of the Management journal, Professor Aleksandar Marković, PhD of linguistic and terminological analyzes. The corpus, however, needs to be expanded for more elaborate scientific research.

Genre. In functional and stylistic sense, the aligned parallel corpus for the domain of management entirely belongs to a single textual genre, i.e. research paper genre. This makes the corpus homogeneous, with uniform level of specialization, similar approach, and no significant variation with regard to register and style. Unlike general corpora, in which genre diversity is recommended, single-genre corpora are entirely acceptable and commonly used in terminology and LSP (language for specific purposes) studies. Since research paper genre is uniform (absent of conversational and dialectical lexicon), dense with terminology, informative, logical, and precise, we believe that it is adequate for terminological research.

2.3 Corpus compilation and preparation for analysis

The process of corpus compilation and preparation for analysis through the use of appropriate software tools consisted of several phases: 1) text preparation and extraction, 2) text alignment at paragraph and sentence level, 3) creation of documents in TEI/XML and TMX formats, 4) metadata supply and 5) insertion into the database.

Text preparation and extraction Upon the selection of texts for the management domain-specific and genre-specific parallelized corpus, all the texts were individually downloaded in PDF format from the journal Management's website. All the texts were then converted into plain text format (.txt) using the *Abby PDF Transformer* program, since this format is standard for corpus processing and analysis software. Each text was renamed for easier identification (e.g. file name Mng52_01-sr refers to the first paper in the 52th issue of the journal in Serbian). During the process, we occasionally encountered the following problems: the converted TXT documents would sometimes lack certain characters and symbols originally present in the PDF format, primarily diacritical markings of the Serbian texts (all the texts in the Serbian language are in Latin script), or two columns from PDF documents would merge into one when converted in plain text format. To minimize these errors, individual files were in some cases first converted into Microsoft Word format (.doc), then corrected, and finally saved as plain text.

After the text conversion, all the elements irrelevant for linguistic analysis (e.g.tables, graphs, charts, formulas, references, contents, headers, footers, etc.) were removed from the corpus.

The texts thus prepared are suitable for the linguistic analysis of non-annotated ("raw") corpora in a language using some of the publicly available corpus analysis programs such as WordSmith[9] or AntConc. [10] The alignment of text at one of its structural levels (section, paragraph, sentence or word level), however, is necessary if we wish to conduct analyses of parallel corpora. In addition, morphosyntactic analysis of the Serbian part of this corpus was also essential for performing an adequate analysis of a highly flective language such as Serbian.

Alignment of texts at paragraph level After the preparation of individual texts, pairs of corresponding Serbian texts and their English translations were aligned at paragraph level (e.g. Mng52_01-en and Mng52_01-en). This process completed using *Notepad ++*, by comparing the contents of paragraph pairs and aligning them, with the aim of having the corresponding paragraphs of Serbian and English text in the same line, each in its own file. We encountered many problems during this process, e.g. untranslated, inadequately translated, missing or misplaced paragraphs or their parts. These issues were solved by finding the missing paragraphs in the original PDF documents, or by removing paragraphs or paragraph parts with no translation equivalents in the other language. The main reason behind this demanding and lengthy procedure is the reduction of noise during future corpus analyses.

Alignment of texts at sentence level and creation of XML documents The third step was the creation of an XML (eXtensible Markup Language) document aligned at sentence level. In addition to texts themselves, XML format texts can also contain additional interpretive linguistic data, i.e. information on text structure, authors, text versions, and the linguistic annotation of the text, including tokenization processes, boundary recognition, morphological analysis (lemmatization and word annotation, part-of-speech / PoS tagging) and shallow parsing.

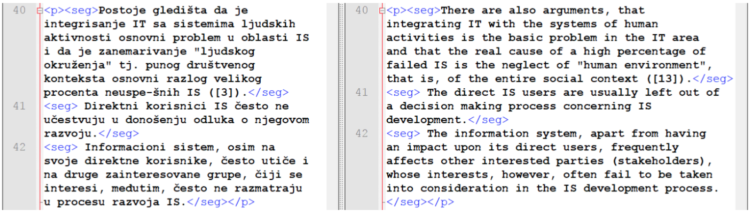
Before the sentence level alignment, texts were segmented into sentences. This step was performed automatically with Unitex (Paumier, 2002), a program that is also used for corpus creation and search. The sentences were segmented using local grammars, i.e. formalisms to describe and recognize

[9] WordSmith (on-line)
[10] AntConc (on-line)

linguistic phenomena in the text. Local grammars were implemented as infinite state automata and transductors and transducers that are manipulated using their graphical representation, i.e. graphs. Local grammars for end-of-sentence recognition are adapted to Serbian language orthography and are an integral part of Serbian language resources, distributed with Unitex. The result of sentence segmentation, i.e. the output text, contains the {S} symbol as the sentence boundary, this is further converted into corresponding TEI / XML format labels for marking sentences (segments) in accordance with the TEI P5[11] Guidelines, the most commonly used unofficial text coding standard.

Text markup at the structural level of paragraph or sentence facilitates the process of pairing source texts with target texts.

In this study, the structural text levels are marked (Figure 1) with labels <div> (entire document), <body> (header), <p> (paragraphs) and <seg> (sentences).



```
<p><seg>Postoje gledišta da je
integrisanje IT sa sistemima ljudskih
aktivnosti osnovni problem u oblasti IS
i da je zanemarivanje "ljudskog
okruženja" tj. punog društvenog
konteksta osnovni razlog velikog
procenta neuspe-šnih IS ([3]).</seg>
<seg> Direktni korisnici IS često ne
učestvuju u donošenju odluka o njegovom
razvoju.</seg>
<seg> Informacioni sistem, osim na
svoje direktne korisnike, često utiče i
na druge zainteresovane grupe, čiji se
interesi, međutim, često ne razmatraju
u procesu razvoja IS.</seg></p>
```

```
<p><seg>There are also arguments, that
integrating IT with the systems of human
activities is the basic problem in the IT area
and that the real cause of a high percentage of
failed IS is the neglect of "human environment",
that is, of the entire social context ([13]).</seg>
<seg> The direct IS users are usually left out of
a decision making process concerning IS
development.</seg>
<seg> The information system, apart from having
an impact upon its direct users, frequently
affects other interested parties (stakeholders),
whose interests, however, often fail to be taken
into consideration in the IS development process.
</seg></p>
```

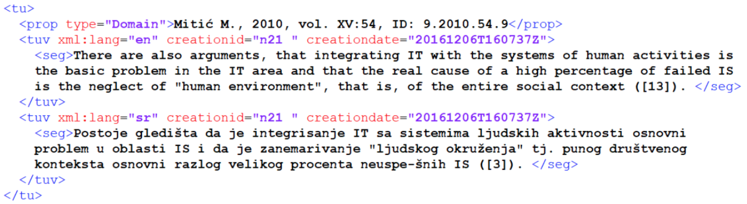
Figure 1. An example of a prepared parallel English-language text in the TEI/XML format

.

Creation of TMX documents The next step was to create TMX format documents (Savourel, 2004). TMX is an XML specification for translation memory data exchange (Translation Memory eXchange) that is often used in computer-aided translation (CAT) tools. The program used for creating TMX documents is ACIDE, an integrated environment for parallelized corpora preparation developed by the Society for Language Resources and Technologies (JeRTeh) in Belgrade (Obradović et al., 2008, 563). ACIDE

[11] TEI P5 (on-line)

offers a graphical interface for alignment and visualization of aligned texts, while the alignment itself is done by XAlign and Concordancier software packages developed in the LORIA [12] laboratory in France (Bonhomme et al., 2001). An example of a paired sentence in TMX format is shown in Figure 2.



```
<tu>
  <prop type="Domain">Mitić M., 2010, vol. XV:54, ID: 9.2010.54.9</prop>
  <tuv xml:lang="en" creationid="n21 " creationdate="20161206T160737Z">
    <seg>There are also arguments, that integrating IT with the systems of human activities is
    the basic problem in the IT area and that the real cause of a high percentage of failed IS
    is the neglect of "human environment", that is, of the entire social context ([13]). </seg>
  </tuv>
  <tuv xml:lang="sr" creationid="n21 " creationdate="20161206T160737Z">
    <seg>Postoje gledišta da je integrisanje IT sa sistemima ljudskih aktivnosti osnovni
    problem u oblasti IS i da je zanemarivanje "ljudskog okruženja" tj. punog društvenog
    konteksta osnovni razlog velikog procenta neuspe-šnih IS ([3]). </seg>
  </tuv>
</tu>
```

Figure 2. An example of a translation unit with an English and the corresponding Serbian sentence in the TMX format

.

The TMX texts were eventually incorporated into a database within MongoDB platform using *Biblisha*; this made the texts available for further search and analysis.

2.4 Supplying the corpus with metadata and inclusion in the database

The prepared TMX documents were incorporated in *Biblisha* (Stanković et al., 2016a) as the seventh collection of aligned parallel texts. The collection itself is divided into 17 sub-collections corresponding to the 17 issues of the Management journal. Each sub-collection contains between nine and 12 documents, i.e. research papers. Each sub-collection and each article have their own unique identification number. For example, identification number 7.2011.59.1 refers to the first article in the 59^{th} issue of the 7^{th} collection, published in 2011. Each article is supplied with bibliographic metadata (in English and Serbian) related to titles, authors, their affiliations and contacts (email addresses), hyperlinks to articles in PDF format, abstracts and keywords in both languages, as well as identification number metadata (article number, issue, year of publication).

[12] LORIA (on-line)

3 Corpus analysis and potential uses

Bilingual (or aligned parallel) domain corpora can be used not only for interlingual comparative and contrastive research, but also for the analysis of its separate segments, i.e. in one or each of the two languages separately. Using one-word and multi-word term extraction and the extraction of translation equivalent pairs in the two languages, we will outline some of the potential uses of our domain corpus.

3.1 Extraction of Serbian keywords

After the texts had been compiled and processed, we followed the keyness criterion to extract Serbian lexical units that are significantly more frequent in the domain corpus than in the reference corpus. The reference corpus used for this purpose was Modern Serbian Language Corpus SrpKor 2003, with 122 million words,[13] created at the Faculty of Mathematics, University of Belgrade (Utvić, 2011, 36a-47a). This process was performed with LeXimir, a lexical resource development and management tool developed by JeRTeh, the Society for Language Resources and Technologies (Stanković et al., 2011, 77-84). In the set of 500 extracted key lexical units, there are 327 nouns, 133 adjectives, 36 verbs and 4 adverbs (Figure 3).

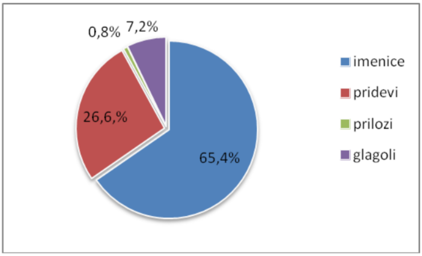


Figure 3. Word class distribution in the set of 500 extracted key lexical units

.

By analyzing and filtering the list of extracted lexical units, we selected the key nouns (Table 2), adjectives (Table 3) and verbs (Table 4). The pa-

[13] SrpKor 2003 (on-line)

rameter of *keyness* is calculated as the ratio of the relative frequency (expressed in millionth parts of the whole, i.e. in ppm as units of measure) in the domain corpus and the corresponding relative frequencies in the reference corpus, with both numbers increased by one before dividing them. Symbols RFr and RFd are relative frequencies (in ppm) in the reference and the domain corpus, while AFr and Afd are the corresponding absolute frequencies. So, the keyness ranks lemmas according to the ratio of frequencies in the domain and the reference corpus, and not only to the frequencies in the domain corpus. The lemmas that appear more frequently in the domain corpus than in the reference corpus, taking into account the corpora size, will be at the top of the table and are most likely to be terms of the management domain.

lemma	keyness	RFr	RFd	AFr	Afd
индикатор	121,398	3,035	488,841	67	299
менаџмент	115,894	16,713	2051,824	369	1255
рачуноводство	115,656	3,306	497,015	73	304
бренд	107,933	1,857	307,365	41	188
портфолио	84,509	1,314	194,555	29	119
интернет	82,145	4,167	423,444	92	259
профитабилност	81,684	1,314	188,016	29	115
перформанса	81,194	4,892	477,396	108	292
подсистем	79,44	0,906	150,413	20	92
сертификација	69,345	1,042	140,603	23	86
конкурентност	67,896	4,529	374,397	100	229
евалуација	65,029	0,725	111,175	16	68
управљање	63,609	38,726	2525,95	855	1545
преференције	62,063	0,544	94,825	12	58
методологија	59,253	6,522	444,698	144	272

Table 2. Key nouns in the corpus

Tables 2, 3 and 4 show that key terminological units do not have to be the most frequent ones. Key terminological units that are also highly frequent in the domain corpus are the most significant terminological units for this domain: the terms *menadžment* (keyness = 115,894, Afd = 1255) and *upravljanje* (keyness = 63,609, Afd = 1545). These synonymous use of these

lemma	keyness	RFr	RFd	AFr	Afd
пројектни	215,922	3,578	987,491	79	604
корпоративан	146,506	2,31	483,936	51	296
одржив	123,302	3,397	541,158	75	331
мотивациони	95,619	0,498	142,238	11	87
ефективан	85,99	2,491	299,19	55	183
рачуноводствени	84,553	2,763	317,174	61	194
екстерни	83,349	2,582	297,555	57	182
иновативан	83,031	1,178	179,841	26	110
управљачки	76,955	4,303	407,095	95	249
стратегијски	68,548	5,979	477,396	132	292
организациони	68,428	20,518	1471,427	453	900
проблемски	64,179	2,582	228,889	57	140
конкурентски	62,603	5,571	410,365	123	251
менаџерски	61,658	2,808	233,793	62	143

Table 3. Key adjectives in the corpus

lemma	keyness	RFr	RFd	AFr	Afd
фокусирати	47,821	3,397	209,27	75	128
имплементирати	40,691	2,038	122,619	45	75
генерисати	33,435	2,355	111,175	52	68
израчунавати	31,611	1,359	73,571	30	45
класификовати	20,779	2,038	62,127	45	38
базирати	19,884	10,644	230,524	235	141
рангирати	19,208	2,627	68,667	58	42
позиционирати	17,702	0,996	34,333	22	21
дефинисати	17,287	53,628	943,348	1184	577
формализовати	17,149	0,679	27,794	15	17
операционализовати	15,316	0,453	21,254	10	13
обухватати	15,087	36,778	568,952	812	348
креирати	15,048	14,494	232,159	320	142
инкорпорирати	15,039	1,132	31,063	25	19

Table 4. Key verbs in the corpus

two terms, however, is often called into question (to be discussed in more detail in Section 3.3). As opposed to this, verbs such as *operacionalizovati* (keyness = 15,316, Afd = 13), *inkorporirati* (keyness = 15,039, Afd = 19) and *pozicionirati* (keyness = 17,702, Afd = 21), for example, have a relatively high key parameter, but they are not highly frequent in the domain corpus.

3.2 Multi-word term extraction in Serbian

Since most terminological units are multi-word (Krstev et al., 2015), the extracted list of one-word terminology units is not sufficient for a terminological analysis. For this reason, we have chosen to include multi-word units automatically extracted from the domain corpus using syntax graphs developed within the Unitex program. By using the tool Leximir, term candidates were extracted according to pre-defined syntactic patterns (Stanković et al., 2016b)(Krstev et al., 2015); the lexical units were then lemmatized to unify all the occurrences of multi-word lemmas. Since this kind of lemmatization can lead to ambiguity, we employed different strategies to resolve these issues. Firstly, lists of lemmas were generated for each term candidate. Secondly, several statistical measures were implemented to rank the term candidates. Finally, term candidate were evaluated and selected for a terminology dictionary (Stanković et al., 2016b). The 20 most frequent Serbian multi-word terminological units in the corpus (shown in Table 5) are mostly noun phrases with an adjective as a premodifier (grf01, the adjective + noun pattern, marked with AXN, with adjective- noun agreement in gender, number and case), e.g. *ljudski resursi, informacioni sistem, elektronsko poslovanje, upravljačko računovodstvo*, etc. In addition, the majority of terminological units shown in Table 5 are multidisciplinary; i.e. not management domain-specific, but rather common for a number of related domains (e.g. *upravni odbor, kamatna stopa, ekonomska kriza, finasijsko sredstvo*, etc.).

The extraction and thorough analysis of terminological units from the Serbian part of our domain corpus can be of great importance not only for the study of semantic, pragmatic and sociolinguistic aspects of management terminology and contrastive and comparative terminology studies, but also contribute to terminological language policy, planning, systematization and standardization of terminology in the domain of management. Even though the examples shown above primarily relate to its applications in terminology studies, our further research will indicate its potential use in the studies of the Serbian language for specific (management) purposes, specialized management discourse, academic writing, genre-analysis, etc.

Graph	Pattern	lemma	frequency	word no.	relative frequency
grf01	AXN	људски ресурси	258	2	421.81
grf01	AXN	организациона наука	207	2	338.43
grf01	AXN	информациони систем	190	2	310.63
grf01	AXN	управни одбор	181	2	295.92
grf01	AXN	електронско пословање	162	2	264.86
grf01	AXN	пројектно финансирање	136	2	222.35
grf01	AXN	пројектни менаџмент	128	2	209.27
grf01	AXN	конкурентска предност	127	2	207.63
grf01	AXN	управљачко рачуноводство	122	2	199.46
grf01	AXN	пројектни менаџер	100	2	163.49
grf01	AXN	финансијски извештај	98	2	160.22
grf03	N2X	реализација пројекта	97	2	158.59
grf03	N2X	управљање ризиком	97	2	158.59
grf01	AXN	информациона технологија	95	2	155.32
grf01	AXN	каматна стопа	95	2	155.32
grf01	AXN	енергетска ефикасност	93	2	152.05
grf03	N2X	процес управљања	92	2	150.41
grf10	2XAXN	јавно-приватно партнерство	90	3	147.14
grf01	AXN	економска криза	86	2	140.6
grf01	AXN	финансијско средство	77	2	125.89

Table 5. The most frequent multi-word terminological units in the domain corpus

3.3 Extraction of English translation equivalents

The previous two sections outline the basic results of analyzing the Serbian part of this corpus. *Biblisha* allows its registered users to access full texts in the corpus via `http://jerteh.rs/biblisha/` website. Without registration and authorization, only its use is limited to the first nine sentences of each text and 30 concordances.

By entering a query (i.e. a one- or a multi-word terminological unit), either in English or in Serbian, into the Biblisha search bar, we obtain English - Serbian concordance pairs. These pairs do not only provide us with a translation equivalent of the entered query, but also with the context of its use in both languages.

Metadata	Concordances (En)	Concordances (Ser): 1260
Milićević et al., 2009, vol. XIV:53, ID: 7.2009.53.1	The modern business prefers an integrative approach in the implementation of **management** tools in tracking . . .	U savremenom biznisu je poželjan integrativni pristup korišćenju **menadžerskih** alata u praćenju . . .
Milosavljević et al., 2012, vol. XVII:62, ID: 7.2012.62.5	As an integral function of global **management**, human resource **management** has a task to explore and define, . . . what it is that makes the organization unique on the market.	**Menadžment** ljudskih resursa, kao integralna funkcija globalnog **menadžmenta**, ima zadatak da istražuje i definiše, . . . šta je to što organizaciju čini jedinstvenom na tržištu.
Đurić D., 2010, vol. XIV:55, ID: 7.2010.55.2	n40 Essentially, financial accounting is part of **management** accounting and reporting.	n40 Suštinski, finansijsko računovodstvo predstavlja deo **upravljačkog** računovodstva i izveštavanja.
Mason R., 2012, No. 63, ID: 7.2012.63.1	The executives in our studies, even those that recognize the opportunities of this still emerging environment, still have widely divergent views on the most effective **management** models for realizing these opportunities.	Direktori obuhvaćeni našom studijom, čak i oni koji su svesni mogućnosti koje daje ovo novo okruženje koje je još u nastajanju, još uvek se razlikuju u stavovima o tome koji su najefektivniji modeli za realizaciju ovih mogućnosti.
Pinterić U., 2008, vol. XIII:49/50, ID: 7.2008.49-50.7	n78 Slovenian scientists wrote about introducing new public **management** elements into the work at all levels of Slovenian public administration . . .	н78 n78 Slovenački autori pisali su o uvođenju elemenata nove javne **uprave** u poslovanje na svim nivoima slovenačke javne uprave . . .
Stanić S., 2008, vol. XIII:49/50, ID: 7.2008.49-50.6	n75 The internal factors include: the amount of media budget, the competence of **management** and administrative structure within the media department of the company or the hired marketing agency.	Interni faktori obuhvataju: veličinu medijskog budžeta, sposobnosti **rukovodeće** i administrativne strukture u okviru medijskog odelenja kompanije ili angažovane marketing agencije.
Domazet et al., 2009, vol. XIV:51, ID: 7.2009.51.4	n160 The Customer Relationship **Management** combines the business strategy and technology aiming to identify, attract and retain long-term relations with customers . . .	н160 Customer Reationship Management kombinuje poslovnu strategiju i tehnologiju sa ciljem da identifikuje, privuče i održi dugoročne odnose sakupcima . . .
Vulić et al., 2012, No. 63, ID: 7.2012.63.7	The **management** is making organizational improvements in the country.	**Rukovodstvo** se organizaciono usavršava u zemlji.
Hitka et al., 2009, vol. XIV:51, ID: 7.2009.51.8	n11 Managers from the area of manpower **management** have to deal with . . . the problem . . .	n11 Menadžeri koji **upravljaju** ljudskom radnom snagom moraju da nađu pravi odgovor na pitanje . . .
Panić S., 2012, No. 63, ID: 7.2012.63.9	To facilitate a two-way communication between the **management** and the employees, the company implemented three modes of communication. . .	Da bi olakšala dvosmernu komunikaciju između **uprave** i zaposlenih, kompanija je uvela 3 načina komunikacije. . .
Barjaktarović et al., 2011, vol. XVI:61, ID: 7.2011.61.1	A very important constituent of the overall bank **management** process is the implementation of the corporate governance principles.	Vrlo bitan element celokupnog procesa **upravljanja** bankom jeste i prime-na principa korporativnog upravljanja.

Table 6. Concordances of the English term *management* and its parallel concordances in Serbian

By typing in the English term *management* into the search bar, for example, we obtain both its concordances in English (with the queried term marked in blue color), and aligned parallel sentences (translations) in the Serbian language. These can be further used to extract Serbian language equivalents of the queried English term (Table 6). In addition to the morphological expansion of the query, Biblisha also enables us to expand the query semantically by using WordNet semantic network and several termbases. The system can find equivalents in the other language, thus enabling the extraction of aligned parallel sentences that find equivalents in 1) both languages,

2) only in Serbian, or 3) only in English. This enables users to exploit the system in various ways.

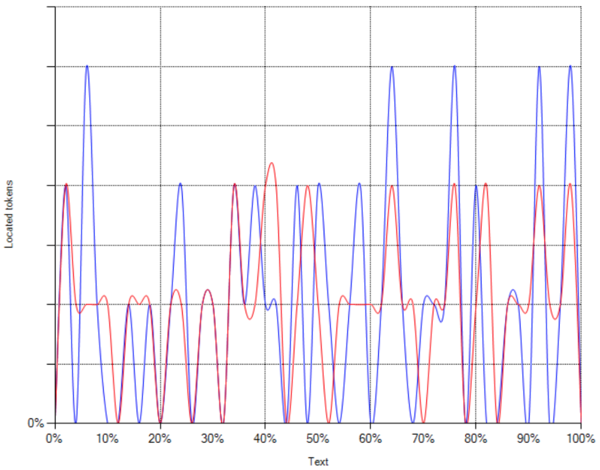


Figure 4. Distribution of the Serbian *менаџмент* и the English *management* in the corpus

.

Table 6 primarily points to different translation equivalents of the English term management in the Serbian language, such as *menadžment, upravljanje, rukovođenje, uprava, upravljački, rukovodeći, menadžerski*, etc., but also to examples of transferring the English term into Serbian without translating it. The Serbian equivalents of *management* extracted from the corpus indicate that this is a polysemous term (management as a process or as a group of people), but also that is has synonyms (e.g. *menadžment, upravljanje* and *rukovođenje* (management as a process), or *menadžment, uprava* and (management as a team of people). Additionally, a more detailed analysis would identify the context in which this English term is translated as an adjective, a verb, or otherwise; this will, however, be discussed in another paper.

A more detailed query, i.e. the search of pairs that consist of the English term *management* and each one of its translation equivalents in Serbian separately (e.g. *management* and *upravljanje*, *management* and *menadžment*, *management* and *rukovođenje*) can provide us with examples of concordances that illustrate why this term is translated in a certain way in the given context. Figure 4 shows the diachronic distribution of the English *management* and the Serbian *menadžment* throughout our domain corpus, from the first papers published in 2008 (far left) to the last ones in the corpus of published 2012 (far right).

A review of *management* translation equivalents in Serbian and the number of corresponding concordances shown in Table 7 indicate that the results of the query contain diverse flective forms in Serbian, and not only the lemma (the nominative case form). The possibilities of using an aligned parallel management corpus illustrated above suit the needs of technical translators, researchers in the fields of comparative and contrasting linguistics and terminology, teachers and students of English for specific purposes, and other user profiles.

Firstly, the use of Biblisa for corpus search can help expert interpreters solve terminological and other language issues in the translation process and find an appropriate translation equivalent in the context of language use, especially since there is a lack of sufficiently available and adequate terminographic and lexicographic resources in the Serbian language.

Secondly, this corpus is a useful resource for comparative and contrastive studies of Serbian and English for specific purposes, e.g. in contrasting the characteristics of academic writing in the two languages, in the terminology variation (e.g. synonymy) studies, and the studies of term usage inconsistencies and gaps that inevitably occur in Serbian as it is a passive recipient of scientific, technological and knowledge transfer coming from developed (mostly English-speaking) countries.

Thirdly, the pedagogical use of aligned parallel corpora is relatively new area of research, explored by, for example, Danielsson and Mahlberg (2003), Granger (1998) and, for Serbian, Ristović (2012).

Monolingual corpora, however, have been used in foreign language teaching and material design. Our aligned corpus parallel presented in this paper can be applied in teaching both directly and indirectly. Indirectly, the English part of the corpus can be used as a basis for the creation of teaching materials, tests and curriculum design for English language courses aimed at management and organization students or professionals. In the indirect corpus use, teachers can create so-called *lexical silabi* (McEnery and Xiao, 2011) by using lists of frequent and key words and expressions as a starting point. The direct exploitation of the corpus refers to *data-driven learning*, a process in which students use the corpus independently (Römer, 2011).

In other words, students of English for management purposes would, with teacher supervision and adequate training in corpus use and exploitation of Biblisha, be able to use the corpus independently in order to explore grammatical, lexical, discursive and other rules and characteristics, or to do error analysis in academic writing (mostly made due to the mother tongue interference), but also in the translation process.

English *management*		example		
Translation equivalents	No.of concordances	source	English	Serbian
управљање	1431	Barjaktarović et al., 2011, vol. XVI:61, ID: 7.2011.61.1	A very important constituent of the overall bank **management** process is the implementation of the corporate governance principles.	Врло битан елемент целокупног процеса **управљања** банком јесте и примена принципа корпоративног **управљања**.
менаџмент	1089	Mitrić et al., 2012, No. 65, ID: 7.2012.65.5	The fields of her scientific and professional interests are related to Accounting and Finance.	Њени главни истраживачки и наставни интереси везани су за област рачуноводства и финансијског **менаџмента**.
руковођење	14	Michalski G., 2008, vol. XIII:49/50, ID: 7.2008.49-50.12	**n95** Operating cycle management should also contribute to realization of this fundamental aim.	**н95** Постизању овог основног циља треба да допринесе и руковођење пословним циклусом.
управа	26	Savoiu et al., 2008, vol. XIII:49/50, ID: 7.2008.49-50.1	**n16** 55 Development of Slovenian selfgovernment in the new public **management** perspective	**н16** 55 Развој локалне самоуправе у Словенији у светлу Нове јавне управе
управљачки	83	Petrović S., 2009, vol. XIV:51, ID: 7.2009.51.6	**n120** • Organizations are too complicated to be understood by means of one **management** model...	**н120** • Организације су превише компликоване да би могле бити схваћене коришћењем једног **управљачког** модела...
руководећи	2	Petković M., 2009, vol. XIV:51, ID: 7.2009.51.1	**n11** ... that are presented by the number and the density of communications among organizational parts, **management** positions or members of a team.	н11 које се представљају бројем и густином комуникација између делова организације, **руководећих** позиција или чланова једног тима.
менаџерски	33	Petković et al., 2012, No. 64, ID: 7.2012.64.7	... organizational design is a **management** lever (tool) used to achieve a balance between effectiveness and efficiency...	...организациони дизајн **менаџерска** полуга (алат), којом се балансира између ефективности и ефикасности...

Table 7. Translation equivalents of the English term management in examples from the corpus

4 Conclusion

The greatest value of the the aligned parallel corpus for the domain of management presented above lies in the fact that prior to its creation there were no other electronically available, aligned and annotated Serbian language corpora for the domain of management or related disciplines (economics, marketing, organization, etc.). Another one of its values is that it can be continually upgraded with new material, thus remaining relevant and up-to-date.

Aligned parallel domain corpora are primary resources for terminology extraction and the production of secondary terminological resources - bilingual terminology dictionaries and termbases, as well as their continuous upgrade with new terms. Such corpora are also useful in the fields of statistical and neural machine translation. As a translation resource, a parallel corpus that is aligned at sentence level can be used to create translation memories and thus facilitate the translation process. Although this paper focuses on Serbian terminology, this resource can also be used for bilingual term extraction.

In addition to the above mentioned applications of aligned parallel corpus for the management domain, there are numerous other possibilities of its application in other types of linguistic research, primarily in terminological, comparative and contrastive linguistic research, translation studies, teaching and learning English as a foreign language, semantic, pragmatic and sociolinguistic studies.

References

Bonhomme, P, TMH Nguyen and S O’ROURKE. “XAlign: l’aligneur de Langue & Dialogue, 2001”

Danielsson, P and M Mahlberg. “There is more to knowing a language than knowing its words: Using parallel texts in the bilingual classroom”. *English for Specific Purposes World. Online Journal for Teachers* Vol. 3, no. 6 (2003)

Erjavec, Tomaž and Nancy Ide. “The MULTEXT-East Corpus”. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 971–74. Citeseer, 1998,

Erjavec, Tomaž, Ana-Maria Barbu, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík et al. “MULTEXT-East "1984" annotated corpus 4.0”, (2010) Fang, Cheng-yu. “Building a corpus of the English of computer science”. *English Language Corpora: Design, Analysis and Exploitation. Amsterdam and Atlanta, GA: Rodopi* (1993): 73–8

Flowerdew, Lynne. “The argument for using English specialized corpora to understand academic and professional language”. *Discourse in the professions: Perspectives from corpus linguistics* (2004): 11–33

Gledhill, Chris. “The discourse function of collocation in research article introductions”. *English for Specific Purposes* Vol. 19, no. 2 (2000): 115–135

Granger, Sylviane. “The computer learner corpus: A testbed for electronic EFL tools”. *Linguistic databases* (1998): 175–88

Krstev, Cvetana, Ranka Stankovic, Ivan Obradovic and Biljana Lazic. “Terminology Acquisition and Description Using Lexical Resources and Local Grammars.”. In *TIA*, 81–89. 2015

Krstev C., Vitas D. “Konkordancije paralelizovanih tekstova”. *Zbornik radova XXXVIII konferencije ETRAN, Niš, juni 1994*, 229–230. 1994

Ljubešić, Nikola, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička and Antonio Toral. “Serbian-English parallel corpus srenWaC 1.0”, 2016

McEnery, Anthony M. and Anita Wilson. *Corpus linguistics: an introduction.* Edinburgh University Press, 2001

McEnery, Tony and Richard Xiao. “What corpora can offer in language teaching and learning”. *Handbook of research in second language teaching and learning* Vol. 2 (2011): 364–380

McEnery, Tony, Richard Xiao and Yukio Tono. *Corpus-based language studies: An advanced resource book.* Taylor & Francis, 2006

Obradović, I, R Stanković and M Utvić. “An integrated environment for development of parallel corpora”. *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen* (2008): 563–578

Paumier, Sébastien. “Manuel d’utilisation du logiciel Unitex”. *Université de Marne-la-Vallée*, 2002

Pazienza, Maria Teresa, Marco Pennacchiotti and Fabio Massimo Zanzotto. “Terminology extraction: an analysis of linguistic and statistical approaches”. In *Knowledge mining*, 255–279. Springer, 2005,

Pearson, Jennifer. *Terms in context*, Vol. 1. John Benjamins Publishing, 1998

Ristović, Zoran. “From corpus to classroom: The use of aligned corpora in English language teaching”. *Infoteka* Vol. 13, no. 2 (2012): 52–66

Roe, Peter Joseph. “The Notion of Difficulty in Scientific Text”. PhD. thesis, University of Birmingham, 1977

Römer, Ute. “Corpus research applications in second language teaching”. *Annual review of applied linguistics* Vol. 31 (2011): 205–225

Savourel, Y. “TMX 1.4 b Specification, The Localisation Industry Standards Association (LISA)”, 2004

Siddiqi, Sifatullah and Aditi Sharan. “Keyword and keyphrase extraction techniques: a literature review”. *International Journal of Computer Applications* Vol. 109, no. 2 (2015)

Sinclair, John. “Corpus and Text: Basic Principles. Wynne, M.(Ed.) Developing Linguistic Corpora: A Guide to Good Practice: 1-16”, , 2005

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

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić and Aleksandra Trtovac. “Rule-based automatic multi-word term extraction and lemmatization”. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC*, 507–514. 2016a

Stanković, Ranka, Cvetana Krstev, Duško Vitas, Nikola Vulović and Olivera Kitanović. “Keyword-based search on bilingual digital libraries”. In *Semanitic Keyword-based Search on Structured Data Sources*, 112–123. Springer, 2016b

Tognini-Bonelli, Elena. *Corpus linguistics at work*, Vol. 6. John Benjamins Publishing, 2001

Utvić, Miloš. “Annotating the corpus of contemporary Serbian”. In *Proceedings of the INFOtheca ‘12 Conference*, 2011

Véronis, Jean. *Parallel Text Processing: Alignment and use of translation corpora* Vol. 13. Springer Science & Business Media, 2013

Vitas, Duško and Cvetana Krstev. “Electronic edition of Serbian translation of Orwell’s 1884 aligned with 7 languages by Duško Vitas, Cvetana Krstev”. 1998a

Vitas, Duško, Goran Nenadić and Cvetana Krstev. “Electronic edition of Serbian translation of Plato’s Republic aligned with 17 languages by Duško Vitas, Goran Nenadić, Cvetana Krstev”. 1998b

Serbian Language Integration in Prolexbase Multilingual Dictionary

UDC 811.163.41'322.2
DOI 10.18485/infotheca.2018.18.2.2

ABSTRACT: In this paper we present the multilingual dictionary of proper names *Prolexbase*, particularly the Serbian volume. We also present the complexity of proper names in the Serbian language, especially those related to their translation: their orthography, derivation, inflection and dialect variations. We describe the model of the *Prolexbase*, with the emphasize on the solutions that had to be made to include the Serbian language in the database (the use of two alphabets, several levels of derivations, the existence of multiple forms). At the end, we give some figures that corroborate the presence of the Serbian language in the *Prolexbase*.

KEYWORDS: proper names, multilingual database, ontology of proper names, LMF format, Serbian language, Prolexbase.

PAPER SUBMITTED: 4 December 2018
PAPER ACCEPTED: 14 December 2018

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

Denis Maurel
denis.maurel@univ-tours.fr
University of Tours, France

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

1 Motivation

As other particularities of language (neologisms, multiwords, idioms and so on), proper names can be responsible of amazing errors. For instance, how should *Bush* be translated to Serbian: as *грм* (plant) or *Буш* (personal name)? Are *Casablanca* and *White House* the same location? It is a common belief that proper names cannot be translated. In fact, all the sorts of translation processes (adaptation, layer, literal translation and so on) are used by translators when they transfer them from a source-language text to a target-language text (Lecuit et al., 2011).

Proper names are also a challenge for Natural Language Processing (NLP) and, more generally, for *Named Entity* tasks[1]. First tasks related to

[1] Named Entities are usually defined by a referent or a kind of uniqueness.

named entities were to complete data bases in the Message Understanding Conferences MUC-6 and MUC-7 conferences with answers to the questions, such as "who did a terrorist attack", "where?", "when?" or "what firm took holdings in another one?", "at what height?", "for how much dollars?" and so on (Chinchor, 1997). Today, the challenge is almost the opposite: entities in a text have to be linked with database entries (Hachey et al., 2013), i.e. proper names have to be disambiguated (see for instance the Text Analysis Conferences (McNamee et al., 2010)). One often uses for these tasks Wikipedia and a number of other semantic data bases, as DBpedia (Auer and Lehmann, 2007), GeoNames, YAGO2 (Hoffart et al., 2012), BabelNet (Navigli and Ponzetto, 2012). These databases constitute a part of the Link Open Data system (LOD) where proper names have a particularly important place.

Prolexbase is a Multilingual Relational Database of Proper Names (Maurel, 2008). The aim of Prolexbase is to assist in their translation. It merges morphology, derivation and semantic relations. For instance, if a sentence in Serbian *Београдска жена ми је рекла да је Дунав прелеп* has to be translated, it may be helpful to expand it: *The female [inflection] inhabitant of the city [semantic expansion] of Belgrade [derivative relation] in Serbia [accessibility relation] has told me that the Danube River [semantic expansion] is splendid.* We will return to this example at the end of this paper.

The first version of Prolexbase (covering eight languages, French, German, English, Italian, Dutch, Polish, Portuguese and Spanish) has been supported by the French *RNTL-Technolangue Project* (2003-2005). Actually, the model of the database was constructed and its coverage for French was good while for other languages it was weak. However, at the same time, an *Egide Pavle Savic* project (2004-2005) has initiated with the aim to add the Serbian language to Prolexbase. The involvement of the Serbian team was very important as it helped to realize the complexity of the morphology and the derivation in the model, that was too French/English centered. Another problem was the presence of two scripts, Cyrillic and Latin. In this first version, a not satisfactory solution was chosen for the problem of two scripts: two volumes were built for the Serbian language, one using the Cyrillic alphabet and the other using the Latin alphabet.

The second version of Prolexbase has been supported by the *Hubert Curien Polonium* project which brought a good coverage for Polish and English Savary et al. (2013). The Serbian part has been significantly improved in the third version of Prolexbase, as a result of a one month visit of Professor Cvetana Krstev which was sponsored by The University of Tours. We improved the coverage of Serbian and we mostly mixed the two alphabets representations in only one volume. We also prepared a possible description of dialect forms, as *Ekavian* and *Ijekavian*.

2 Prolexbase

2.1 The Prolexbase model

Since Prolexbase is a multilingual databases we need a model enabling the linking of different occurrences of proper names in different languages. We choose to define the linguistic class of proper names (and their derivations) as an ontology in the sense of (Gruber, 1995): "A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose… An ontology is an explicit specification of a conceptualization".

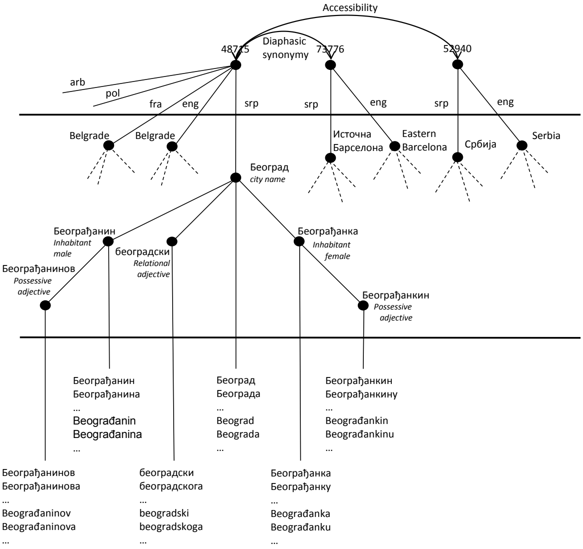


Figure 1. The example of *Београд* in the Prolexbase model

The center of the Prolexbase ontology is the *conceptual proper name*, the *pivot*, which represents the referent from a particular point of view. For instance, *Pope Francis* and *Jorge Mario Bergoglio* or *Београд* (*Belgrade*) and *Источна Барселона* (*Eastern Barcelona*). The translation by pivots (Boitet, 1988) is not very usual today, although a pivot can be refined for some language and not for others. For instance, in the Papillon project (Mangeot, 2000), the pivot for *rice* in English corresponds to two refined pivots in Japanese, *raw rice* and *cooked rice*. For the *conceptual proper name*, no refinement is needed, so we can use this model without any problems. For each language, the pivot is linked to an unique set of proper names, the *prolexeme*. This set contains the proper name and, eventually, its aliases and its morphosyntactic derivatives (see 2.3). The pivots constitute the conceptual level of the model and the prolexemes its linguistic level. The ontology is completed by two other levels, one at the top, the metaconceptual level (types and supertypes), and the other one at the bottom, the instance level (forms of proper names – as they appear in a written text). Figure 1 illustrates the model with the proper name *Београд*.

2.2 The LMF format

Prolexbase is a free and open source resource, under LGPL-LR license[2]. The exchange format is inspired by the Lexical Markup Format (LMF) (ISO/TC 37/SC 4, 2007). Figure 2 shows LMF classes for representing the Prolexbase model. It represents a selection of classes of the LMF core model with additional parts from LMF extensions (packages Morphology, NLP semantics, NLP multilingual notations and NLP syntax). Multilingual descriptions are represented with grey boxes. The whole resource is represented by the class *Lexical Resource* to which some information is linked, such as the language codes, scripts, characters used in the whole resource (class *Global Information*). The resource contains the conceptual level (class *Sense Axis*) and the linguistic level, with several lexicons (class *Lexicon*) that are monolingual descriptions. One of them is the Serbian lexicon. The lexical entries are all lemmas of a prolexeme (proper names, aliases and derivatives) with their word forms (all the instances): classes *Lexical Entry*, *Lemma*, *Word Form* and *Form Representation*.

These lemmas are linked with the senses that are pivots with category assigned, as *relational adjective* or *male possessive adjective* (classes *Sense* and *Sense Relation*). These pivots are defined in the class *Senses Axis* belonging to the multilingual part of the resource. The class *Sense Axis Relation*

[2] Prolexbase (on-line)

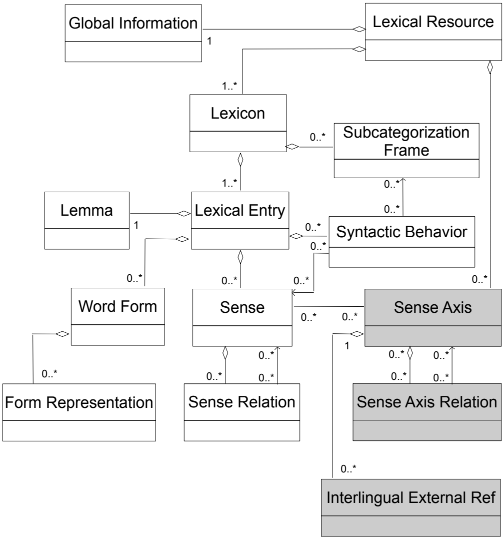


Figure 2. The LMF schema of Prolexbase

represents relations between conceptual proper names while the class *Interlingual External Ref* represents typologies. We also note some information about classifying contexts of proper names (class *Subcategorization Frame* and some idiosyncratic collocations (class *Syntactic Behavior*). These classes are not yet used in the Serbian volume and some examples would be the use of different prepositions for proper names, like in *Србија је* ***на*** *Балкану и* ***у*** *Европи* (*Serbia is in the Balkans and in Europe*).

2.3 Relations

The large part of Prolexbase consists of relations between pivots (language independent relations): synonymy, meronymy and accessibility.

The synonymy relation, or more precisely, the quasi-synonymy relation, is a relation between two pivots referring to the same referent for which different points of view exist. The translator has to choose the same point of view, which is not always possible. We distinguish between three different points of view, designated by three diasystematic features of (Coseriu, 1998):

- Diachronic: variations depending on time. *Савезна Република Југославија* (*Federal Republic of Yugoslavia*) versus *Државна Заједница Србија и Црна Гора* (*the State Union of Serbia and Montenegro*);
- Diastratic: variations depending on sociocultural stratification. *Јосип Броз* (*Josip Broz*) versus *Тито* (*Tito*);
- Diaphasic: variations depending on the usage purpose. *Београд* (*Belgrade*) versus *Источна Барселона* (*Eastern Barcelona*).

The meronymy relation, a partitive relation, is a relation of inclusion. The examples are geographical inclusion: Serbia is in the Balkans, which is a part of Europe, and temporal relation: The bombing of Belgrade on April 6, 1941 is a part of (happened during) the Second World War. We extend this relation to other domains, such as economy, nationality and so on.

The accessibility relation (Ariel, 1990), an associative relation, means that a proper name is accessible through some other proper names. In dictionaries, proper names, contrary to common nouns, do not have definitions – they are rather replaced by a relation to some more known name. Thus, this relation is rarely symmetric: in dictionaries, one can read that Aaron is the brother of Moses, but Moses is not presented as the brother of Aaron, but as the leader of the Hebrews. Consequently, Aaron is accessible through Moses and Moses is accessible through the Hebrews' story. We distinguish 12 such relations:

- Relative: *Арон* (Aaron) is the brother of *Мојсије* (Moses);
- Capital: *Београд* (Belgrade) is the capital of *Србија* (Serbia);
- Leader: *Тито* (Tito) is a political leader of *Југославија* (Yugoslavia);
- Founder: *Растко Немањић* (Rastko Nemanjić) founded the *Српска православна црква* (Serbian Orthodox Church);
- Follower: *Петар* (Peter) is a disciple of *Исус* (Jesus);
- Creator: *Госпођа министарка* (The Cabinet Minister's Wife) is a comedy by *Бранислав Нушић* (Branislav Nušić);
- Manager: *Ранко Жеравица* (Ranko Žeravica) was a Serbian basketball coach who used to manage the *Југословенска кошаркашка репрезентација* (Yugoslav national basketball team);

- Tenant: *Александар Вучић* (Aleksandar Vučić) is the tenant of the *Нови двор* (Novi dvor);
- Heir: *Кнез Михаило Обреновић* (Prince Mihailo Obrenović) was the heir of *Кнез Милош Обреновић* (Prince Miloš Obrenović);
- Headquarters: In *Београд* (Belgrade) are the corporate headquarters of *Montinvest Beograd*;
- Rival: *Партизан* (Partizan) is the football rival of *Црвена звезда* (Red Star);
- Companion: *Мирко* (Mirko) was *Славко*'s (Slavko) beloved comrade and brother-in-arms.[3]

The language dependent relations are frequency, is-an-alias, is-a-derivative, collocation, context and eponymy.

A prolexeme is the set of all lemmas semantically linked to a proper name in one particular language. For instance, the prolexeme *Београд* (*Belgrade*) consists of: Београд, београдски, Београђанин, Београђанка, Београђанинов, Београђанкин as shown in Figure 1. The three main relations at the language dependent level are frequency, is-an-alias and is-a-derivative. The frequency which indicates whether the proper name is well known can have three possible values: commonly used, infrequently used and rarely used. Today, this frequency can be calculated from LOD, mainly from Wikipedia (Elashter and Maurel, 2016). Aliases are different variations of a proper name: short forms, abbreviations, acronyms, different orthographies, alternate transcriptions, diatopic quasi-synonyms and explanations. Only the morphosemantic derivatives of a prolexemes are considered (and their derivatives). For instance, *пастеризовати* (to pasteurize), referring to the process of partial sterilization, is a derivative of the name *Пастер* (Pasteur), but it is not semantically linked to it.

Collocation and context relations concern the local usage of proper names. In some languages, such as French, country names are often preceded by an article, masculine or feminine without any particular reason which one should be used: for instance, one says *la* (feminine) *France* and *le* (masculine) *Montenegro*. The context is a relation between a proper name and typical words appearing with it. The context can be classified as a classifying or an accessibility context. The classifying context is an expansion of a noun phrase (capital, king, coach, etc.), called by MacDonald (1990) the *external structure* of a proper name. The classifying context is can be

[3] Two main characters of a very popular Yugoslav comic book series about two Partisan couriers.

useful in translation. For instance, *Сава* (*Sava*) is translated in English as the *Sava River*. The accessibility context is a noun phrase that implements the accessibility relation between two pivots. It can be regarded as a sort of the explanation of a proper name by a link to a well-known proper name. For instance, one can translate *Београд* (*Beograd*) as *Belgrade, the capital of Serbia*.

The eponymy relations differs from other relations: it tells us that the translation does not refer to a proper name but to a common noun (antonomasia), as *жилет* (*žilet*) in Serbian that designates all razor blades and not only the Gillette ones, or to a terminological term, as *Parkinson's disease* or *Pythagoras' theorem*, or again to an idiomatic phrase, as *све ми је равно до Косова* (*sve mi je ravno do Kosova*) that says literally *It's all straight to me up to Kosovo* and idiomatically *It's all equal to me* or *I don't care at all.*

2.4 Typologies

The meta-conceptual level deals with the concept existence and the typology of proper names.

The existence concept divides proper names into three groups: the historical ones that exist or existed, as *Београд* (*Belgrade*); the religious ones whose existence depends on one's beliefs, as *Арханrел Михаил* (*Michael the Archangel*), or fictional ones invented by authors. Generally, name belonging to two later categories need to be translated, as *Snow White* that is translated in Serbian as *Снежана*.

The aim of Prolexbase typology is to classify proper names. We defined four big classes (named *supertypes*) corresponding to the primary semantic features: the human (*Anthroponyms*), the location (*Toponyms*), the concrete (*Ergonyms* – artifacts and work names) and the event (*Pragmonyms*). We defined thirty types in total, presented in Table 1. This typology defines the primary hypernymy relation between a pivot and type. We completed it with another relation, the secondary hypernymy relation, which is a metonymy relation between types, as seen in Table 2.

3 Proper names in the Serbian language

3.1 Alphabets

In Serbia, the use of Cyrillic alphabet is prescribed by law (Zakon, 2010, article 1), while the use of Latin alphabet is permitted in special situations

Proper Name						
Anthroponym			Ergonym	Pragmonym	Toponym	
Individual	Collective					
		Group				Territory
Celebrity	Dynasty	Association	Object	Disaster	Astronym	Country
Patronymic	Ethnonym	Ensemble	Work	Feast	Building	Region
First name		Firm	Thought	History	Geonym	Supranational
Pseudo-anthroponym		Institution	Product	Manifestation	Hydronym	
		Organization	Vessel	Meteorology	City	
					Way	

Table 1. The Prolexbase typology – the primary hypernymy

(traffic signs, street names, etc.). However, due to historical and other reasons Latin alphabet is widely used and it is defined as equal to Cyrillic in the Serbian orthography (Пешикан et al., 1993, articles 1–6). The Serbian alphabet, both Cyrillic and Latin, has 30 letters; 1-1 correspondence is established between these two sets, as presented in Table 3. The order of letters in Cyrillic and Latin alphabet is different; letters in Table 3 are presented in the Cyrillic order. The Serbian Latin alphabet does not use some of the 26 letters of English alphabet[4] – Q, W, X and Y. It uses some letters with diacritics – Č, Ć, Đ, Š and Ž – while some are represented as combinations of existing letters as digraphs – Lj, Nj and Dž. Digraphs are in electronic texts traditionally represented with two codes of consisting letters, although Unicode has introduced specific codes for these symbols[5]. One should note that capital letters of digraphs lj, nj and dž can be represented in two ways, with only the first composing capital letter – Lj, Nj and Dž – and with both capital letters – LJ, NJ and DŽ. The later case is used when the whole word (or a longer text) is written in capital letters. This is reflected in Unicode as well that has separate codes for these representations.

3.2 Names of foreign origin

Proper names of foreign origin, as a rule, are not written in Serbian using the original script and spelling, they are rather transcribed. It is applied

[4] Letters represented in ASCII code

[5] See code page Unicode Latin Extended-B

Types	**Secondary hypernymy**
Country Region Supranational Territory	Collective anthroponym
City	Collective anthroponym Ergonym
Buiding Way	Ergonym
Feast History Manifestation	
Association Ensemble Firm Group Institution Organization	Ergonym Toponym
Vessel	Toponym

Table 2. The secondary hypernymy relation

to personal names as well as geopolitical names. The Orthography manual (Пешикан et al., 1993, articles 101–180) permits the use of the original script and spelling for Serbian texts written in Latin; however, in practice it is rarely used. One of the reasons is that the use of transcription for both scripts facilitates publications in both of them, as well as switching between them on Web pages[6].

In Serbian, orthographic, or practical, transcription is used to customize sounds from the original language to the standard Serbian spelling system. The Orthography manual (Пешикан et al., 1993, articles 101–180) lists transcription rules for 27 languages, including Latin, Ancient and Modern Greek, Japanese and Chinese. However, there are number of proper names that do

[6] For instance, all articles in Serbian Wikipedia can be viewed in Cyrillic and Latin alphabet – see, for instance, the Wikipedia page about *Orthographic transcription* in Serbian. The same possibility is offered by some newspaper portals, for instance *Politika*.

Cyrillic	а А	ђ Ђ	ј Ј	н Н	с С	х Х
Latin	a A	đ Đ	j J	n N	s S	h H
Cyrillic	б Б	е Е	к К	њ Њ	т Т	ц Ц
Latin	b B	e E	k K	nj Nj	t T	c C
Cyrillic	в В	ж Ж	л Л	о О	ћ Ћ	ч Ч
Latin	v V	ž Ž	l L	o O	ć Ć	č Č
Cyrillic	г Г	з З	љ Љ	п П	у У	џ Џ
Latin	g G	z Z	lj Lj	p P	u U	dž Dž
Cyrillic	д Д	и И	м М	р Р	ф Ф	ш Ш
Latin	d D	i I	m M	r R	f F	š Š

Table 3. The Serbian alphabet – Cyrillic and Latin; the order of Cyrillic alphabet is present top-down, from left to right.

not conform to these rules, mostly because they are used as such for a very long time, or because they better suit Serbian language and its morphological properties. Some examples listed in the Orthography manual for geographic names are *Москва* (Moscow) (instead of *Масква*), *Волгоград* (*Volgograd*) (instead of *Валгаграт*), *Њујорк* (*New-York*) (instead of *Њујок*) and *Лајпциг* (*Leipzig*) (instead of *Лајпцих*) and for personal names *Ганди* (*Gandhi*) (instead of *Гандхи*) and *Strindberg* (instead of *Стриндберј*). For some foreign geographic names the Serbian name is neither the original nor its transcription, for instance *Беч* for *Vienna*.

Multi-unit geographic names are, as a rule, transcribed into multiword names, for instance, *Њу Хемпшир* (*New Hampshire*) and *Солт Лејк Сити* (*Salt Lake City*). There are exceptions to this rule as well, for instance *Порторико* (*Puerto Rico*). The foreign geographic multi-unit names that have as constituents one or more common words are sometimes translated, partially translated or not translated at all. For instance, *Rocky Mountains* is translated as *Стеновите планине*, while *Long Island* is transcribed as *Лонг Ајланд*. The same common words are sometimes translated and sometimes transcribed, for instance *Нови Јужни Велс* (*New South Wales*) vs. *Њу Делхи* (*New Delhi*).

Sometimes multiple variants exist in Serbian for a single foreign proper name. For instance, *Кот д'Ивоар* is the transcribed name in official use for *Côte d'Ivoire* while its translated name *Обала слоноваче* prevails in everyday use. It is more often the case for names of location with mixed inhabitants, like *Целовец* and *Клагенфурт* (Klagenfurt) (a city in Austria), or for names of locations that changed names due to political reasons, for instance *Град Хо Ши Мин* (*Ho Chi Minh City*), former *Сајгон* (*Saigon*).

Transcription rules are not always easy to master, so additional manuals are published that can help in writing proper names of foreign origin, for instance *The transcription dictionary of English personal names* (Prćić, 1992) and *The English-Serbian dictionary of geographic names* (Prćić, 2004). It is unfortunate that information in these manuals sometimes contradicts the Orthography manual: for instance, in (Prćić, 2004) the transcription for *Rio de Janeiro* is *Рио де Жанејро* while the Orthography manual for the same name suggests *Рио де Жанеиро* that does not conform completely to the transcription rules for Portuguese but is established as a name.

Organization names are specific as compared to other proper names. They are more often than others used in original, especially acronyms such as *IBM* or *FBI*. Besides that, organization names can be transcribed *Мајкрософт* (Microsoft) or translated *Организација за економску сарадњу и развој* (Organization for Economic Cooperation and Development). Moreover, for some organizations both transcribed and translated names are used, for instance *Британска телевизијска мрежа* and (rare) *Бритиш броудкастинг корпорејшн* (British Broadcasting Corporation). The corresponding acronyms can be either in original *BBC* or in spelling *Би-Би-Си* (Krstev et al., 2015).

3.3 The derivation

Nouns and adjectives can be derived from most geographic proper nouns[7].

Names of inhabitants, or demonyms, are derived from various geographic proper names: continents, super-regions, countries, regions, cities, and city districts as represented in Table 4[8]. For some names of these types it is not possible to derive a name for an inhabitant, e.g. *Осло* (*Oslo*) and a phrase is used instead *становник Осла* (an *inhabitant of Oslo*). If a name of a male inhabitant can be derived, than, as a rule, a name for a female inhabitant can also be derived, and for both of them possessive adjectives can be

[7] We will not consider verbs derived from geographic proper nouns, such as *пофранцузити се* (*become as a Frenchman/Frenchwoman*) as explained in Subsection 2.3.

[8] Дорћол (Dorćol) is a central district of Belgrade.

	Name	**Inhabitant (m.)** **Possessive adj.**	**Inhabitant (f.)** **Possesive adj.**	**Adjective**
continent Europe	**Европа**	Европљанин Европљанинов	Европљанка Европљанкин	европски
super-region Balkan	**Балкан**	Балканац Балканчев	Балканка Балканкин	балкански
country France	**Француска**	Француз Французов	Францускиња Францускињин	француски
region Provence	**Прованса**	Провансалац Провансалчев	Провансалка Провансалкин	провансалски
city Belgrade	**Београд**	Београђанин Београђанинов	Београђанка Београђанкин	београдски
city district Dorćol	**Дорћол**	Дорћолац Дорћолчев	Дорћолка Дорћолкин	дорћолски

Table 4. Names of inhabitants and adjectives derived from certain types of toponyms in Serbian

derived, as well as an adjective describing something as related to the initial toponym. For *Осло* such an adjective cannot be derived either. On the other hand, for some names of inhabitants other adjectives can be derived, for instance, *Парижанин* (a *male inhabitant of Paris*) → *Парижанинов* (*belonging to a male inhabitant of Paris*) → *парижански* (*referring to, in a style of inhabitants of Paris*) (opposed to *париски* (*referring to Paris*)). Occasionally, diminutives can be derived from names of inhabitants, e.g. *Српче* and *Српчић*, diminutives of *Србин* (an inhabitant of Serbia), sometimes referring to children.

In certain cases, double or even triple names of male inhabitants can be derived, leading to multiple names of female inhabitants, possessive and descriptive adjectives. The examples are:

- double names derived from *Кореја* (*Korea*)[9]:
    - *Корејац* (m), *Корејка* (f), *Корејчев* (m poss.), *Корејкин* (f poss.), *корејски* (adj.);
    - *Кореанац* (m), *Кореанка* (f), *Кореанчев* (m poss.), *Кореанкин* (f poss.), *кореански* (adj.);

[9] These examples are corroborated in (Стијовић, 2016).

– triple names derived from *Париз* (*Paris*)[10]:
  - *Парижанин* (m), *Парижанка* (f), *Парижанинов* (m poss.), *Парижанкин* (f poss.), *париски* and *паришки* (adj.);
  - *Парижлија* (m), *Парижлијка* (f), *Парижлијин* (m poss.), *Парижлијкин*(f poss.);
  - *Паризлија* (m), *Паризлијка* (f), *Паризлијин* (m poss.), *Паризлијкин* (f poss.).

For certain multi-unit geographic names demonyms and corresponding adjectives are derived either by composing its constituents or by using just one of them. In either case as a result simple words are derived as illustrated in Table 5[11]. However, for a number of multiword geographic names demonyms and corresponding adjectives cannot be derived.

Name – Serbian original	**Inhabitant – male female**	**Adjective**
Кабо Верде (Cabo Verde)	Кабоверђанин Кабоверђанка	кабовердски
Буркина Фасо (Burkina Faso)	Буркинац Буркинка	буркински
Тринидад и Тобаго (Trinidad and Tobago)	становник Тринидада и Тобага становница Тринидада и Тобага	*descriptive*
Нови Сад	Новосађанин Новосађанка	новосадски
Бачко Ново Село	становник Бачког Новог Села становница Бачког Новог Села	*descriptive*

Table 5. Names of inhabitants and adjectives derived from multi-unit names of toponyms in Serbian

Adjectives are derived from other types of geographic names, hydronyms and oronyms, as well. Examples are, hydronyms *дунавски* derived from *Дунав* (*Danube*), *сенски* derived from *Сена* (*La Seine*) and oronyms *алпски*

[10] These examples are corroborated in (Стевановић, 1967).

[11] These examples are corroborated in (Стијовић, 2016).

derived from *Алпи* (*Alps*), *копаонички* derived from *Копаоник* (a mountain in Serbia). For some hydronyms and oronyms adjectives cannot be derived, for instance *Волга* (Volga River). Adjectives derived from multiword hydronyms and oronyms, if they exist, are simple words, for instance, *великоморавски* derived from *Велика Морава* (a river in Serbia) and *старопланински* derived from *Стара Планина* (a mountain in Serbia).

Possessive adjectives can be derived from personal names: first names, surnames and nicknames. For instance, possessive adjectives derived from all parts of the name *Иво Лола Рибар*[12] would be *Ивов*, *Лолин* and *Рибаров.* Various nouns and adjectives can be derived from names of famous persons. For instance, the name of the philosopher *Карл Маркс* (*Karl Marx*) yields in Serbian: *марксизам* for a doctrine, *марксиста* and *марксисткиња* for supporters of *марксизам*, *марксологија* for a scientific discipline, *марксолог* and *марксолошкиња* for scientists studying *марксологија*, the adjectives *марксистички* (*relating to Marxists*) and *марксолошки* (*relating to Marxologs*). Many of these derived adjectives and nouns can be prefixed, eg. with *анти-*, *нео-*, *пост-*, etc. (Vitas and Krstev, 2013). These derivatives are not considered in the Prolexbase, as explained in Section 2.3.

Possessive adjectives can be derived from some, mostly simple-word, organization names. For instance, *Мајкрософтов* is a possessive adjective derived from *Мајкрософт* (*Microsoft*). Possessive adjectives are also used for acronyms of organization names – in such cases, derivational suffixes are added to an acronym after a hyphen, e.g. *IBM-ов* (*belonging to IBM*).

3.4 Grammatical features

Proper names in Serbian, as well as nouns and adjectives derived from them, share inflectional properties with common nounds and adjectives.

The gender of geographic names, toponyms, oronyms and hydronyms, can be masculine, feminine or neuter while the names inflect in cases (seven different cases). They do not inflect in number which can be either singular or plural. The examples are given in Table 6.

Geographic names are, as a rule, inanimate although there are some confusing examples: there are a few city names in Serbia named after some famous people, for instance *Јаша Томић* and *Алекса Шантић*[13]. If they

[12] Ivo Lola Ribar (1916-1943), Yugoslav national hero.

[13] Jaša Tomić (1856 – 1922) was a politician, Aleksa Šantić (1868 – 1924) was a poet.

are considered inanimate, the sentence *I travel to Jaša Tomić* would be in Serbian *Путујем у Јаша Томић* that looks incorrect, since Јаша Томић, as a person is animate[14].

type	**original or English**	**name**	**number**	**gender**
toponym	Belgrade	Београд	singular	masculine
oronym	Olympos	Олимп		
hydronym	Danube	Дунав		
toponym	Karlovci	Карловци	plural	
oronym	Alpes	Алпи		
hydronym	Dardanelles	Дарданели		
toponym	Athens	Атина	singular	feminine
oronym	Aconcagua	Аконкагва		
hydronym	Seine	Сена		
toponym	Budějovice	Будјеовице	plural	
oronym	Divčibare	Дивчибаре		
hydronym	Plitvice	Плитвице		
toponym	Valjevo	Ваљево	singular	neuter
oronym	Pohorje	Похорје		
hydronym	Oranjerivier	Орање		
toponym	Kaštela	Каштела	plural	

Table 6. Geograhic names in Serbian with different gender and number

Demonyms derived from geographic names have masculine gender (for male inhabitants) and feminine gender (for female inhabitants) and they inflect in case and number. Adjectives derived from geographic names inflect

[14] In Serbian, the form of the accusative case singular for the masculine gender nouns depends on the animatness: for the inanimate nouns it is equal to the nominative case while for the animate nouns it is equal to the genitive case. In this example, the preposition *u* invokes the accusative case which for the (inanimate) *Јаша Томић* is the same as the nominative case, while for the (animate) *Јаша Томић* is *Јашу Томића* (for example in the sentence, *Милица се заљубила у Јашу Томића* (Milica fell in love with Jaša Tomić)).

in case, number, gender and animatness. It should be noted that possessive adjectives do not have comparative and superlative forms, and neither do descriptive adjectives except occasionally, e.g. *Војводина је најевропскији део Србије* (*Vojvodina is the most European part of Serbia*).

Serbian first names and nick names can have the masculine or the feminine gender, they are in singular and they inflect in case. Serbian surnames have masculine gender and they, in general, inflect both in number and in case. Some surnames, mostly of foreign origin, do not inflect in number because of morphological restrictions. Complex agreement rules apply to Serbian full names that depend on the gender of a first name and the order a first and a surname in a full name – one rule is that surnames do not inflect in case for female personal names (Gucul-Milojević, 2010). Women are sometimes referred by possessive adjectives of a surname in the feminine gender or by a feminine gender noun derived from a surname by gender motion. Some examples are given in Table 7.

Form	**Surname**	**Feminine forms**
nominative singular	*Петровић*	*Петровићка* *Петровићева*
genitive singular	*Петровића*	*Петровићке* *Петровићеве*
nominative plural	*Петровићи*	*Петровићке* *Петровићеве*

Form	**Full name (male)**	**Full name (female)**
nominative singular	*Петар Петровић*	*Зорка Петровић* *Зорка Петровићка* *Зорка Петровићева*
genitive singular	*Петра Петровића*	*Зорке Петровић* *Зорке Петровићке* *Зорке Петровићеве*

Table 7. Male and female personal names and their inflection

Organization names inflect in case while their number and gender do not change and, in general, depend on organization name form, if a single-word, or on the number and the gender of its head word, if a multiword. For instance, *Мајкрософт* (*Microsoft*) has the masculine gender, while *Сорбона* (*Sorbonne*) has the feminine gender. Among multiword organization names ***Универзитет*** *у Београду* (*University of Belgrade*) has the masculine gender, *Београдска аутобуска* ***станица*** (*Belgrade bus station*) has the feminine gender, while ***Удружење*** *спортских новинара Београда* (*Association of sport journalists of Belgrade*) has the neuter gender. Organization names ***Лекари*** *без граница* (*Doctors Without Borders*) and *Међународне мировне* ***снаге*** (*International peacekeeping forces*) have the plural number[15].

3.5 Dialects

In Serbian, two standard variants of pronunciation are in use, Ekavian and Ijekavian. They differ in the reflection of the old Proto-Slavic phoneme (*jat*): in Ekavian variant it is replaced predominantly by *e*, while in the Ijekavian variant its is replaced by syllables *ije/je*.

These variants do not have big influence on proper names, because most of proper names do not contain the reflection of the phoneme (*jat*). In cases when they do, the name is usually used in one of dialects only. For instance, in two city names *Ријека* (in Croatia) and *Ријека Црнојевића* (in Montenegro) the common noun is used only in Ijekavian dialect – *ријека* (and not *река*) (*river*). On the other hand, the feminine first name derived from the common noun *вера/вјера* (*faith*) – has both the Ekavian *Вера* and the Ijekavian variant *Вјера*. However, such a name in one variant would not change if it appears in a text written in another variant, that is, it is unchangeable.

In organization multiword names various common words appear that can be in either of variants. These variants are then reflected in organization names as well depending on the variant a text in which they appear uses, for instance, Ekavian variant *Светска банка* vs. Ijekavian variant *Свјетска банка* (*World Bank*).

4 Achived Results

4.1 The Serbian language contribution to the Prolexbase model

As we said in section 1, the inclusion of the Serbian language led to the development of a better Prolexbase model. The collaboration between research

[15] Head nouns in these multiword organization names are in bold.

groups from the University of Tours and the University of Belgrade was very fruitful in many respects, but we will emphasize two issues that we consider the most important: the derivation relation and the form representation.

The derivation relation. In Section 3.3 we presented the complexity of derivation rules of the Serbian language, such as the quasi-systematic possibility to create derivatives from human names, and thus from derivatives of topological names as well (the relational or inhabitant names). For instance, (see Figure 1), the city name *Београд* (*Belgrade*) generates (as in many other languages) a derivative *Београђанин* (a male inhabitant of *Belgrade*), while *Београђанин* generates in its turn *Београђанинов* (a possessive adjective of a male inhabitant of *Belgrade*). In English and French only one level of derivation exists: *Belgrade*/*Belgradian* in English and *Belgrade*/*Belgradois* in French. The first database model did not include relation from the table *Derivative* to itself. We added this relation to later models, and then we realized that this relation exists in French as well: for instance, the name of a prize, like Nobel prize, quasi-systematically allows the creation of a verb meaning *to give the prize*, for instance, *nobeliser*, while from such verbs it is possible to regularly create other derivatives, like *nobelisable* (a person who is likely to be chosen by the Nobel Prize Committee), and so on.

The form representation. In the Prolexbase database model, we place proper names in two tables, *Prolexeme* (the longest form of the name) and *Alias* (others forms). However, in the LMF representation (see Figure 2) this distinction disappears, because all aliases are equivalent entries, linked by the sense.

The question was whether a name written in the Cyrillic alphabet and the same name written in the Latin alphabet are aliases or not? However, it would look amazing that the same word can be alias of itself! Sure not. We considered first the possibility to define two prolexemes in Serbian language (Cyrillic and Latin), but we abandoned this idea since this solution violates the constraint of the uniqueness of the pivot projection in a particular language. For that reason we adopted a second solution that defines two lexicons, the Serbian Cyrillic lexicon and the Serbian Latin lexicon. Finally, the third version of Prolexbase was produced at the University of Tours in which we added systematically under the *Word Form* one or more *Form Representations*. For instance, for *Београд* (*Belgrade*) we now have:

```
<LexicalEntry partOfSpeech="noun">
```

```
  <Lemma>Београд</Lemma>
  <WordForm grammaticalGender="masculine"
  grammaticalNumber="singular"
  grammaticalCase="nominative"
  grammaticalAnimacy="nonAnimate">
    <FormRepresentation script="cyrl">
      Београд
    </FormRepresentation>
    <FormRepresentation script="latn">
      Beograd
    </FormRepresentation>
  </WordForm>
  ...
</LexicalEntry>
```

After this choice was done, we added for some entries the distinction between the Ekavian and the Ijekavian dialect (see Section 3.5), for which we used the same LMF representation:

```
<LexicalEntry partOfSpeech="noun">
  <Lemma>Немачка</Lemma>
  <WordForm grammaticalGender="feminine"
  grammaticalNumber="singular"
  grammaticalCase="nominative"
  grammaticalAnimacy="nonAnimate">
    <FormRepresentation script="cyrl">
      Немачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ekavsk">
      Немачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ijekavsk">
      Нјемачка
    </FormRepresentation>
    <FormRepresentation script="cyrl"
    geographicalVariant="ijekavsk">
      Њемачка
    </FormRepresentation>
    <FormRepresentation script="latn">
      Nemačka
    </FormRepresentation>
  </WordForm>
  ...
</LexicalEntry>
```

Finally, we used the concept of *form representation* to some variants of writing, as, for instance, in the example above, *Њемачка* and *Нјемачка* (the later rarely used), but also in some other cases, like the differences in transcription, as *Рио де Жанејро* and *Рио де Жанеиро* (*Rio de Janeiro*) (see Section 3.2), or for different forms for the same set of values of grammatical categories – the surname *Чехов* (*Chekhov*) has three variant singular dative forms: *Чехову*, *Чеховом* and *Чеховому*. We enhanced this approach to other languages as well, eliminating thus the alias category *Variant*.

4.2 The Prolexbase implementation

The Table 8 gives some numbers about the Serbian language implementation. We manually introduced the prolexemes with their link to the pivot from a selection of the French ones, and we also added a few aliases. Then we automatically generated the derivatives, and of course, all the inflections of the prolexemes, aliases and derivatives.

Serbian prolexemes	8 526
Serbian aliases	21
Serbian derivatives	920
Serbian instances	108 325
Serbian pivot relations	29 567

Table 8. Serbian language implementation

We can add to these numbers the amazing[16] number of the instances derived from *Београд*. If we complete the Figure 1 with all instances, we obtain 626 forms...

[16] Compared to English or even French!

Coming back to the example from Section 1: *Београдска жена ми је рекла да је Дунав прелеп* we now obtain:

београдска
  female inhabitant (derivative category)
  Belgrade (prolexeme)
    city (classifying context)
    Serbia (accessibility)
    capital (accessibility context)
→ The female inhabitant of the city of Belgrade, capital of Serbia
жена ми је рекла да је
→ has told me that
Дунав
  river (classifying context)
→ the Danube River
прелеп
→ is splendid

4.3 Conclusion

We showed that the complexity of the Serbian language morphology led to the essential contribution to the Prolex multilingual dictionary project. The necessary improvements that were introduced in order to accommodate Serbian proved to be useful for other languages in the Prolexbase. These improvements particularly concern the treatment of derivations and the representation of multiple forms. This was proved during our work on inclusion of non-European languages, such as Arabic, into the database since its internal structure was able to model them. This work also showed how important it is to include in linguistic multilingual project the variety of languages, not only the proximate ones.

Acknowledgment

The authors thank the University of Tours for enabling this research by funding the Prof. Cvetana Krstev one-month visit to the University. The authors also thank the *Egide Pavle Savic* fund that sponsored the project that initialized the collaboration between universities of Tours and Belgrade in 2004. This research was partly supported by the Ministry of Education, Science and Technological Development through projects 138006 and III47003.

References

Ariel, M. *Accessing Noun Phrases Antecedents*, 1990

Auer, S. and J. Lehmann. “What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content”. In *ESWC 2007*, no. 4519, LNCS, 503–517. 2007

Boitet, C. *Pros and cons of the pivot and transfer approaches to multilingual machine translation*, 93–106. 1988

Chinchor, N. “Muc-7 Named Entity Task Definition”, 1997, URL http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

Coseriu, E. “Le double problème des unitès dia-s”. In *Les Cahiers dia. Etudes sur la diachronie et la variation linguistique, Universitè de Gent, Belgique*, Vol. 1, 9–16. 1998

Elashter, Mouna and Denis Maurel, “Estimer la notoriété d’un nom propre via Wikipedia”, In *TALN 2016. Paris*, 2016, URL https://jep-taln2016.limsi.fr/actes/

Gruber, T. R.. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”. *Int. Journal of Human-Computer Studies* Vol. 43 (1995): 907–928

Gucul-Milojević, Sandra. “Personal Names in Information Extraction”. *INFOtheca* Vol. 11, no. 1 (2010): 53a–63a

Hachey, B, W Radford, J Nothman, M Honnibal and R Curran, J. “Evaluating entity linking with Wikipedia”, In *Artificial Intelligence*, 194, 130–150. 2013

Hoffart, J., F. M. Suchanek, K. Berberich and G. Weikum. “YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia”. *Artificial Intelligence Journal, Special Issue on Artificial Intelligence, Wikipedia and Semi-Structured Resources* (2012)

ISO/TC 37/SC 4. *Language resource management - Lexical markup framework (LMF)*, 2007. http://lirics.loria.fr/documents.html

Krstev, Cvetana, Duško Vitas and Ranka Stanković. “A Lexical Approach to Acronyms and their Definitions”. In *Proceedings of 7th Language & Technology Conference, November 27–29, 2015, Poznań, Poland.* 2015

Lecuit, Émeline, Denis Maurel and Duško Vitas. “A tagged and aligned corpus for the study of Proper Names in translation”. In *Workshop Annotation and exploitation of parallel corpora, International Conference Recent advance in Natural Language Processing (RANLP 2011),*, 11–18. 2011, URL http://aclweb.org/anthology/W11-43

MacDonald, D. *Internal and external evidence in the identification and semantic categorisation of Proper Names*, 21–39. 1990

Mangeot, M. "Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links". In *7th Workshop on Advanced Information Network and System, Kasetsart University, Bangkok, Thailand.* 2000

Maurel, D. "Prolexbase: A Multilingual relational Lexical Database of Proper Names". In *LREC 2008*, 334–338. 2008

McNamee, P., H. T. Dang, H. Simpson, P. Schone and S. M. Strassel. "An evaluation of technologies for knowledge base population". In *LREC 2010*, 369—372. 2010

Navigli, Roberto and Simone Paolo Ponzetto. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". *Artificial Intelligence* Vol. 193 (2012): 217–250

Prćić, Tvrtko. *Transkripcioni rečnik engleskih ličnih imena [Transcription dictionary of English personal names].* Nolit, 1992

Prćić, Tvrtko, *Englesko-srpski rečnik geografskih imena [English-Serbian dictionary of geographic names].* Zmaj, 2004

Savary, A., L. Manicki and M. Baron. "Populating a Multilingual Ontology of Proper Names from Open Sources". *Journal of Language Modelling* Vol. 1, no. 2 (2013)

Vitas, Duško and Cvetana Krstev. "Derivational Morphology in E-Dictionaries of Serbian". In *Proceedings of the 32nd International Conference on Lexis and Grammar, September 10–14, 2013, Faro, Portugal.* 2013

Zakon, eds.. *Zakon o službenoj upotrebi jezika i pisma [Law on Official Usage of Language and Script].* Službeni glasnik Republike Srbije, 2010

Пешикан, Митар, Јован Јерковић and Мато Пижурица, ed.. *Правопис српскога језика [The Orthography of Serbian Language].* Матица српска, 1993

Стевановић, Михаило и др., eds.. *Речник српскохрватскога књижевнога језика [Serbo-Croatian literary language dictionary].* Матица српска, 1967

Стијовић, Рада. "Званични пуни скраћени називи држава на српском и енглеском језику [Official and shorten names of countries in Serbian and English]", 2016, internal report

Open Science Platform — Obligation of Publishing in Open Access in the Republic of Serbia

UDC 027.8(497.11):371.38]:004.738.5
DOI 10.18485/infotheca.2018.18.2.4

Vesna Z. Abadić
vesnaa@kg.ac.rs
Marija M. Gordić
mgordic@kg.ac.rs

University Library
Kragujevac, Serbia

ABSTRACT: By following the work of the university‘s community, one can notice the advantages of the networked digital environment that enables the easier and faster access to published texts. Mass communication that does not violate the code of academic integrity, has led to the emergence of a new paradigm: open science. Library and information systems in Serbia already work on the sustainable development of digital repositories by actively participating in projects supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia. The BE-OPEN project, as an Erasmus+ structural project in the field of capacity building in higher education, gathered the largest state universities. Following the adoption of the Open Science Platform, universities are obliged to create their institutional platforms as a precondition for the application of open science principles.

KEYWORDS: Open science, Open access, Digital repository, Open science platform, BE-OPEN

PAPER SUBMITTED: 15 August 2018
PAPER ACCEPTED: 22 November 2018

1 Open Access Practice in Serbia

Library-information systems of the Serbian universities consist of University libraries and a network of academic libraries. For more than a decade, libraries, as members of the system, have advocated open access publishing and played a key role in maintaining the digital repository infrastructure (zak, 2011). The definition of open access is given by Peter Suber - *Open access literature is digital, online, free of charge, and free of copying and licensing restrictions* (Suber, 2016). Open access provides freedom in exchanging ideas and results of scientific research work to the entire scientific community.

In 2011, as a part of the TEMPUS project "New Library Services at the Universities of the Western Balkans", the universities of Belgrade, Niš and Kragujevac established the PHAIDRA system as a digital repository. The system was taken over from the University of Vienna.[1] Despite of promotions and educations at the mentioned universities, the researchers from the universities did not deposit a great number of facilities. The highest number of digital collections was set by librarians themselves. The more extensive application of PHAIDRA has been obtained by amending the Law on Higher Education,[2] introducing the obligation to deposit defended PhD theses.

Open Science[3]is a principle that promotes and creates free access to scientific knowledge and results of scientific research, without legal, technological or social constraints. Even though the open science is a new term in the academic community, it originates from the end of the 16th and the beginning of the 17th century (David, 2008), when a certain population of people recognized the need for common communication within the same scientific field. This kind of renaissance in science has upstaged the established introvert way of individuals' behavior within the framework of scientific research and began a new "revolution" in academic circles. The result of "sharing" data has led to the increase in "cooperative rivalries" in discovering new knowledge (David, 2008). The results of "sharing" data has led to the increase in "cooperative rivalries" in discovering new knowledge (David, 2004). This social phenomenon has enabled the creation of a large number of easily ac-

[1] PHAIDRA digital repository of the University of Vienna (accessed on 07/26/2018)

[2] By Article 30, paragraph 9 of the Law on Higher Education ("RS Official Gazette", No. 76/2005, 100/2007 - authentic interpretation, 97/2008, 44/2010, 93/2012, 89/2013, 99/2014, 45/2015 - authentic interpretation, 68/2015 and 87/2016) determine that the University is obliged to establish a digital repository in which electronic versions of defended doctoral dissertations are permanently stored, together with the report of the commission for the assessment of the dissertation, the mentor's data and the composition commission and copyright protection data, as well as make all the information available publicly available.

[3] Open science (accessed on 07/26/2018)

cessible data sources and thus facilitated interaction among researchers. The popularity and necessity of monitoring and using all aspects of open science, information access and data sources grow over time (McKiernan, 2016), although it can be said that this concept has not been fully adopted. Today, with the open science, and with the help of electronic communication, we can track all segments of a research process, starting from the methodology of work, through information about the used apparatus and technical means, up to the results of the research. With this sort of "accelerator of knowledge" (Woelfle et al., 2011), the increasing transparency of science is being achieved.

Open science is a wider concept than the open access publishing itself and it is based on several principles:

- Open data obtained from surveys,
- Open source code,
- Open review, scientific communication and methodology,
- Transparent monitoring of the scientific work results for the purpose of further evaluation, using a wide range of indicators, for all types of research results.

In April 2018, the European Commission adopted Open Science Policy Platform Recommendations, outlining eight priorities which encourage its development (Commission, 2017) The emphasis is on open data, their protection and opportunities to be reused (Kanjilal and Das, 2015). It is clear that such a task requires a complex technical infrastructure, compatible with the existing systems and competencies of all participants in the process. The obligation of publishing papers and research results generated in projects funded by the European Union in open access has been adopted by the European Commission in the form of *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research data in Horizon 2020* (H20, 2017). Serbian researchers who participated in international projects have deposited their works on various platforms or networks (Zenodo, Research Gate, etc.), because there was no legal regulation which covers this area.

2 Implementation of the BE-OPEN project

The academic community is aware of the need for publishing in an open access. There are individual examples of good practice of some Serbian scientific institutes which have established institutional or thematic repositories for years. The fact that the Republic of Serbia's budget for scientific development cannot be compared with the budget of EU countries was a key motive for finding and developing additional opportunities in the framework of projects financed by European Commission, as well as for taking established solutions. Since the largest universities in Serbia were aware of the fact that this area has to be formally regulated, they have gathered in BE-OPEN[4] (*Boosting Engagement of Serbian Universities in Open Science*) Erasmus+ project (K2 area - capacity building in higher education). Project partners are six state universities, as well as the Ministry of Education, Science and Technological Development of the Republic of Serbia. The project is implemented from October 2016 to October 2019.

The main goal of the project is development of implementation conditions of full principles of open science, which will be realized through the following steps:

- Development of national and institutional legal acts and guidelines (platforms),
- Implementation of institutional digital repositories at all universities in Serbia, as well as the National Portal of Open Science,
- Strengthening individual competencies by organizing seminars, conferences and workshops for all potential users of the repository,
- Integration through the establishment of the National Open Source Portal, which will provide transfer of technology and knowledge from the academic community to industry and general public, and at the same time provide analytical data for the analysis of research results.

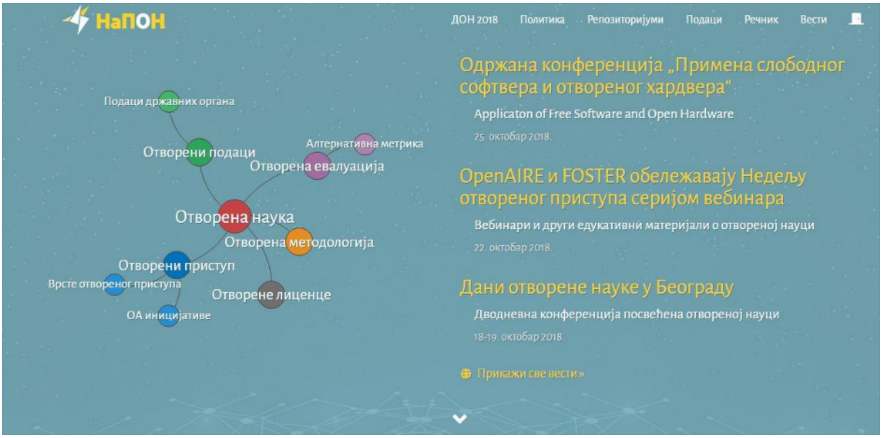
In the first year of project implementation, a detailed analysis of the current situation was provided, the reports on the adopted legislation and the current practice of open science in Serbia were published and placed on the project website, within the WP-1 work package.[5]

At the beginning of November 2017, the website of the Open Science Portal was created, which will be the hub of open science in Serbia, available at `http://www.open.ac.rs/`.[6] NaPON - as the National Open Science Portal, will connect all existing and future digital repositories of individual institutions and data from state authorities. It will provide a starting point for all information in the field of open science.

[4] The front page of the BE-OPEN project. (accessed on 07/28/2018)

[5] WP-1 (accessed on 10/20/2018)

[6] National Portal of Open Science. (accessed 10/28/2018)



Одржана конференција „Примена слободног софтвера и отвореног хардвера“

Applicaton of Free Software and Open Hardware

25. октобар 2018.

OpenAIRE и FOSTER обележавају Недељу отвореног приступа серијом вебинара

Вебинари и други едукативни материјали о отвореној науци

22. октобар 2018.

Дани отворене науке у Београду

Дводневна конференција посвећена отвореној науци

18-19. октобар 2018.

Figure 1. The homepage of the National Open Science Portal

.

3 Open Science Platform

In accordance with its goals, on July 9, 2018, the Ministry of Education, Science and Technological Development of the Republic of Serbia, adopted the Open Science Platform (pla, 2017) The Platform is intended for all participants in scientific research activities and refers to the results of research projects and programs completely or partially funded from the budget of the Republic of Serbia. The realization of the open science principles, which will entail the full protection of ethical standards, copyrights of intellectual property, will take place in four directions:

- An open approach to scientific literature,
- Availability of data collected in scientific research,
- Transparency of scientific communication and methodology, and
- Development of digital infrastructure that will enable the realization of the above-stated goals.

The Ministry prescribes the norm that the integral text of published results should be originally in open access, and at the latest 12 months from the date of publishing in the field of natural, medical and technological sciences, or 18 months for research in social sciences and humanities. Immediately after publishing, metadata of scientific publications should be deposited into an institutional or national repository. The obligation to deposit in repositories will also refer to journal articles, monographs and conference proceedings that are already published in open access. In case that a paper was previously published in a commercial publisher journal (a holder of the copyrights), the terms of the contract signed by the author with the publisher must be respected. It is possible to deposit published version (with the permission of the publisher) and with respect to the embargo period, or the peer-reviewed version that has been accepted for publication. Prior to submitting the manuscripts, it should be useful for authors to check the publishing policy of the chosen journal on the portal SHERPA/RoMEO,[7] which contains unified publishing policies of various journals.

The common scientific publishing practice worldwide is that authors pay the publisher a fee for publishing costs in an open access (*APC - Article Processing Charges or BPC - Book Processing Charges*). In order to meet the needs of an author, it has announced that these costs could be an item in the budget of projects financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

Metadata should always be available to the publicity and unique identifiers of articles and researchers, such as the DOI (*Digital Object Identifier*) and ORCID (*Open Researcher and Contributor ID*), will be standard elements of metadata set. ORCID represents the unique international affiliate-related researcher identifier, which has become an essential element in publishing in numerous foreign journals. A profound and persistent approach of librarians at university and faculty libraries to create complete researcher bibliography, and efforts to link these data to the corresponding ORCID identifiers, will allow the connection of data from a mutual bibliographic database to objects in open access.

Regarding the primary research data, the question of the justification of their archiving was raised. The ability to provide long-term storage in machine-readable formats, and availability within an interoperable digital platform, in order to be used again in other surveys or further research, is a sufficient reason for this venture. After archiving, the degree of data availability is defined. Metadata must always be visible. Future institutional platforms will prescribe the conditions for the deposit of primary data, in accordance with legal or ethical limitations, which will be classified into three groups of availability:

- Closed data,

[7] SHERPA/RoMEO (accessed on 08/02/2018)

- Data available to a defined group of researchers, or
- Fully publicly available.

Adequate treatment of research data consists of several steps, including collection, processing, data analysis by the selected statistical method, defining ways of keeping and allowing reuse by other researchers. Future protection of data will depend on the type of data (for example, whether personal data is included). Data exchange between different research groups in different disciplines can enable the use of such data in new research programs and thus bring savings. Requirements to be findable, accessible, interoperable and reusable are named FAIR principles of research data.

Like the existing university digital repositories in Serbia, the future digital platforms must provide interoperability, which implies the possibility of automatic data downloading in accordance with the international protocol OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting), and the structured metadata scheme in line with the Dublin Core[8] standard. Until now, repositories for the storage of research data have not been developed in Serbia, so it is recommended by the Ministry that researchers can deposit their data in the appropriate international thematic repositories in the absence of institutional resources.

4 Legal provisions

Machine-readable Creative Commons (CC) licenses[9] will be generally applied to every deposited object. CC licences define the conditions under which research results can be further used. For the full implementation of these provisions, a six-month period was adopted, in which universities and institutes, as independent units, must adopt their institutional platforms for open science. As a part of the final provisions of the Open Science Policy Platform, it has been stated that the Ministry will monitor the compliance of all above-mentioned norms, since the final results will influence the future financing of new projects and other activities under the Ministry's competencies. It remains to be seen which way they will be monitored and whether there will be consequences for disrespecting norms. In the framework of its legal acts (institutional platforms), the institutions will have the obligation to define the level of obligation of publishing in open access, which

[8] Dublin Core metadata initiative. (accessed on 08/02/2018)

[9] CC BY 3.0 RS. (accessed 08/02/2018)

can be verified in different ways, for example, in the evaluation procedure for advancement in academic career. Since the level of commitment has been adopted from the adoption moment of the legislation, the question is to what extent the scientific works will be found in the repositories before the legal obligation. Possible automatic downloading of previously published publications, which were earlier deposited in other repositories, with the download of available data from citation databases and other open archives, will be complemented by the open science corpus in Serbia. There are a large number of scientific research works from Serbia that are already available in open access. Those papers are deposited on numerous portals and aggregators around the world. It is not a rare phenomenon that an authors' affiliation is not seen, or it has not been updated. Additional engagement enables the linking of these works with institutional repositories, which ultimately will have an impact on dissemination of Serbian scientific thought as well as the authors' citation and the overall position of each university on one of the ranking scales.

5 Conclusion

From all mentioned above, it can be concluded that implementation of full principles of open science will contribute to better dissemination and visibility of scientific production, its adequate evaluation and greater utilization of the scientific results. Encouraging research processes, increasing the visibility of results with the possibility that they will be used more than once (Swan, 2016), brings benefits to the general public. Systematic monitoring the results of scientific work, respecting copyright and related rights of publicly available data in use, their evaluation in project decision making etc., will be achieved by adopting numerous unified procedures in work. Considering all these aspects for achieving principles of open science, it is clear that a whole range of technical, legal, organizational elements and time are needed in order to notify the full effects of the application. Certainly, the practice of universities and institutes in Serbia is a proof that that certain awareness about the importance of this process exists. The future infrastructure will bring together all actors in the process of creating scientific results: institutions, researchers, publishers, reviewers, librarians, on the one hand, and financiers and beneficiaries of the scientific community results, on the other (ministry, business entities). It will contribute to more efficient connection between science and economy, and science and the entire society through the development of innovative services and products. Librarians have an active role in this process: starting from education and assistance in relation to depositing procedures, metadata control, system maintenance, and up to monitoring the implementation of the Open Source Platform. In accordance with the institutional regulations that will be adopted, the role of librarians in academic institutions is and will be crucial in mediating open science processes.

References

“Закон о библиотечко-информационој делатности”, 2011

“H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020”, 2017, accessed on 07/28/2018, http://www.mpn.gov.rs/wp-content/uploads/2018/07/Platforma-za-otvorenu-nauku.pdf

“Платформа за отворену науку”, 2017. Приступљено 30.7. 2018, http://www.mpn.gov.rs/wp-content/uploads/2018/07/Platforma-za-otvorenu-nauku.pdf

Commission, European. “OSPP-REC: Open Science Policy Platform Recommendations”, 2017, accessed on 07/28/2018, https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf#view=fit&pagemode=none

David, Paul A. “Understanding the emergence of open science institutions: functionalist economics in historical context”. *Industrial and Corporate Change* Vol. 13, no. 4 (2004): 571–589. Приступљено 30. 10. 2018, https://academic.oup.com/icc/article-abstract/13/4/571/718486?redirectedFrom=fulltext

David, Paul A. “The Historical Origins of ’Open Science’: An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution”. *Capitalism and society* Vol. 3, no. 2 (2008). Приступљено 30. 10. 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2209188&gt

Kanjilal, Uma and Anup Kumar Das. “Introduction to open access”, 2015, Преузето 30. 7. 2018, http://unesdoc.unesco.org/images/0023/002319/231920E.pdf

McKiernan, Erin C. “Point of view: How open science helps researchers succeed”, 2016, Приступљено 30. 10. 2018, https://elifesciences.org/articles/16800

Suber, Piter. “Otvoreni pristup”, 2016

Swan, Alma. "Policy guidelines for the development and promotion of open access", 2016, Приступљено 30. 10. 2018, http://unesdoc.unesco.org/images/0021/002158/215863e.pdf

Woelfle, Michael, Todd Olliaro Piero and H. Matthew. "Open science is a research accelerator". *Nature Chemistry* Vol. 3 (2011): 745–748, Приступљено 30. 10. 2018, https://www.nature.com/articles/nchem.1149

Focus on Open Science at the University Library "Svetozar Markovic"

Ivana Gavrilović
ivana.gavrilovic@gmail.com

University Library
"Svetozar Markovic"
Belgrade, Serbia

PAPER SUBMITTED: 22 November 2018
PAPER ACCEPTED: 1 December 2018

Belgrade University Library "Svetozar Markovic" hosted one-day international workshop titled *Focus on Open Science (Chapter VIII)* on November 12, 2018. The workshop is the result of cooperation between the library and Scientific Knowledge Services with the support of LIBER.[1] The workshop is intended for librarians and scientists, and it's the eighth in a series of the same chapters organized around Europe.[2]

After the introductory presentation given by the director of the library, Prof. Aleksandar Jerkov, many relevant topics regarding the Open Science today were addressed during four sessions. The emphasis was primarily on the role of libraries, especially academic ones, in the field of *Open Science*; institutional, regional and international exchange of experiences; necessity of increasing the awareness of the importance of this topic; ways of financing; as well as the advocacy strategies, all for the purpose of applying the best practice.

The moderator of the Belgrade workshop was Dr Tiberius Ignat.[3] He is the CEO of a company whose objective is to help European libraries improve their work through the implementation of new technologies and to maximize their potential for providing, exchanging and using information and digital services. The visitors were encouraged to interactively participate in the workshop, through social networks or directly.

[1] *Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries*

[2] Previous workshops took place in Budapest, Ljubljana, Barcelona, and Gdansk. Next one is in Vienna.

[3] Tiberius Ignat, *Scientific Knowledge Services*

In the first session, Dr Paul Ayris,[4] one of the authorities in this field, and Vanessa Proudman,[5] director of SPARC Europe, opened the discourse. They pointed out the significance of the Open Science concept, its importance in democratization of knowledge and the necessity of making scientific work accessible to wider community. Dr Paul Ayris introduced the European Commission Platform,[6] which consists of a series of recommendations focusing on eight key areas of Open Science: rewards and Incentives, research indicators and next-generation metrics, future of scholarly communication, research integrity, skills and education, citizen science, FAIR[7] data and *European Open Science Cloud*. He also pointed out a new model of publishing papers via *Open Access Press*, which was developed at the London University College, where he works as a director of UCL Library Services. By embracing this form of open publishing, scientific research papers from London University College are now easily and promptly available worldwide. He particularly emphasized the need to raise the awareness in scientific communities, the necessity to change obsolete patterns of thinking and re-examine the policies of the universities themselves.

Vanessa Praudmen pointed out the importance of providing support to researchers and students, in the form of seeking new models of financing and developing new tools for managing scientific data. She also expressed the need for international cooperation, so that the idea of Open Science could be sustained in presence and expanded in future.

During the second session, Dr Ignasi Labastida i Juan,[8] from the University of Barcelona, and a representative of LIBER, Dr Giannis Tsakonas,[9] discussed the topic from the perspective of universities and researchers. Dr Ignasi Labastida i Juan pointed out that the universities have to change

[4] Paul Ayris (University College London, UK); lecture title: *Be the change that you wish to see in the world*

[5] Vanessa Proudman (*Scholarly Publishing and Academic Resources Coalition – SPARC*); lecture title: *It's high time to rethink how we pay for Open Infrastructure; it's high time to act*

[6] *Open Science Policy Platform*

[7] The acronym originates from the following terms: Findability, Accessibility, Interoperability, and Reusability

[8] Ignasi Labastida i Juan (*Universitat de Barcelona, Spain*); lecture title: *Facing the Open Science challenges from a university perspective*

[9] Giannis Tsakonas (*University of Patras, Greece*); lecture title: *A new kind of dialogue : Open Science as a discourse topic between libraries, researches and society*

the way they manage scientific information. When comparing to the past, information is now stored, disseminated, published and used in a different way, and even the course of scientific research is conducted quite differently. Universities should transform their policies in accordance with the principles of Open Science. Giannis Tsakonas presented a strategy called *Open Road Map*. LIBER, as the largest network of research libraries in Europe, has written a new strategy of action and proposed a number of practical steps that libraries should undertake to support Open Science at both local and international levels.

Miro Pušnik,[10] director of Central Technological Library and one of the hosts of the first held workshop in Ljubljana in 2015, shared with audience his experience regarding the implementation of Open Science in Slovenia. He pointed out the role that academic libraries have in transforming the paradigm of the scientific-research process, highlighting the importance of building good infrastructure and educating librarians as the missionaries of these changes.

The central idea presented by Dr Adam Sofronijević,[11] deputy director of the University Library "Svetozar Marković", implies the change in the way contemporary society perceives institutions, primarily scientifically oriented (universities, libraries, scientific centers, consulting agencies, ministries, etc.). The focus of these institutions is shifting from mere interest in science towards business, and they lose their basic role of "guardians of the truth". The access to these institutions is limited and their work is almost oblique to the public. On one hand citizens have lost faith in these institutions, while on the other they have at their disposal huge amounts of information that is often unverified, even inaccurate. Therefore, due to the previously mentioned, evolving the institutions through the concept of Open Science is a must in order to regain the lost trust and steer further development of science in the appropriate direction.

Open science is a movement that is based on openness, cooperation and exchange. It is focused on the development of science for the benefit of humanity. Libraries, especially academic ones, must recognize the importance of their role in transition towards new models of openness in society, and they need to find their rightful place. It's not enough only to have good faith, they also need to understand the rights and obligations they have regarding

[10] Miro Pušnik (*University of Ljubljana, Slovenia*); lecture title: Encouraging Open Science Literacy: *For years of Focus on Open Science workshops*

[11] Adam Sofronijević (*University of Belgrade*); lecture title: *Open Science: The Remedy for a Post-truth World*

scientific work. Continuous education, proficiency in new technologies and cooperation with similar institutions in order to exchange good practice are imperative. Focus on Open Science workshops have an important role in providing guidance and recommendations for this transition by pointing to all the benefits of Open Science: easier and faster access to scientific information, greater transparency of scientific work, economic aspect, giving society a wider access to science and many others. They advise on how the change in society and its institutions should occur, but they also caution against traps and problems that follow the transition to Open Science: copyright restrictions, the issue of evaluating a large number of scientific papers that appear in open access, additional administration for scientists when publishing their papers (traditionally done by journal publishers) and others. The workshop in Belgrade certainly achieved this goal.

Summer School "ESSLLI 2018"

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs
Faculty of Philology
University of Belgrade
Serbia

PAPER SUBMITTED: 5 December 2018
PAPER ACCEPTED: 18 December 2018

1 About Summer School

European Summer School in Logic, Language and Information, ESSLLI 2018, was held 6–17 August 2018. It took place in Sofia, the capital of Bulgaria. Flawless organisation was not surprising, since the school has been held for the thirtieth time this year. The previous schools were held in Toulouse (France), Bolzano (Italy), Barcelona (Spain), Tübingen (Germany), Dusseldorf (Germany), etc. The following institutions organised the event: Sofia University "Sv. Kliment Ohridski", Institute for Information and Communication Technologies, Bulgarian Academy of Sciences, and Bulgarian Association for Computer Linguistics. All the lectures were held at the premises of the University's main building. An official poster of the school can be seen in Figure 1.

2 Organisation

Programme committee was composed of:

- Chairing Laura Kallmeyer (Düsseldorf University, Germany)
- Co-chairing Galia Angelova (IICT-BAS, Sofia, Bulgaria)
- (Language and Computation) Noah Goodman (Stanford University, USA) and Barbara Plank (University of Groningen, The Netherlands)
- (Language and Logic) Márta Abrusán (CNRS, IRIT Toulouse & IJN Paris, France) and Robert Levine (Ohio State University, USA)
- (Logic and Computation) Wojciech Jamroga (Polish Academy of Sciences, Warsaw, Poland)

Organising committee was composed of:

- Petya Osenova (Sofia University, Bulgaria)

Venue:
Sofia University
“St. Kl. Ohridski”

Organized by:
Sofia University “St. Kl. Ohridski”
and
Institute of Information
and Communication
Technologies,
Bulgarian
Academy of
Sciences

Sofia, Bulgaria
August 6-17 2018

30th European Summer School
in Logic, Language and Information

ESSLLI 2018
August 6-17, 2018, Sofia, Bulgaria
http://esslli2018.folli.info/

Figure 1. A poster for the school

- Kiril Simov (IICT-BAS, Sofia, Bulgaria)
- Galia Angelova (IICT-BAS, Sofia, Bulgaria)
- Svetla Boytcheva (IICT-BAS, Sofia, Bulgaria)
- Vladislava Grigorova (IICT-BAS, Sofia, Bulgaria)
- Ivan Koychev (Sofia University, Bulgaria)
- Ivelina Nikolova (IICT-BAS, Sofia, Bulgaria);
- Tsvetomira Pashova (Sofia University, Bulgaria)
- Alexandra Soskova (Sofia University, Bulgaria)
- Irina Temnikova (Sofia University, Bulgaria)
- Anelly Kremenska (Sofia University, Bulgaria)
- Gueorgui Jetchev (Sofia University, Bulgaria)

With a support of around twenty more people, this team was responsible for the impeccable organisation of the ESLLII 2018 school.

3 Courses

The school offers almost 50 courses. Each course lasts seven and a half hours (five working days, an hour and a half per day). Each of the courses is classified into one of the following three categories:

- Language and Logic;
- Logic and Computing;
- Language and Computing.

Courses are also classified by level, and within each category there are fundamental, introductory and advanced courses. All the classes are held in English.[1,2]

A participant can attend eight courses at most, since there are four time slots in five days, both weeks.

09.00–10.30	Lecture
10.30–11.00	Coffee break
11.00–12.30	Lecture
12.30–14.00	Lunch break
14.00–15.30	Lecture
15.30–15.50	Coffee break
15.50–16.50	Student session
17.00–18.30	Lecture
18.30–19.00	Coffee break
19.00–20.00	Evening lecture

In addition to various courses, workshops can also be attended by the participants, that are more practically-oriented. In the first week, the following workshops were organised:

- Ambiguity: Perspectives on Representation and Resolution
- Bridging Formal and Conceptual Semantics
- Annotation in Digital Humanities (annDH): How Can Linguistics/Computational Linguistics Help with Annotation in DH

During the second week, the following workshops were held:

[1] Review of the first week (on-line)

[2] Review of the second week (on-line)

– NLP in the Era of Big Data, Deep Learning, and Post Truth
– Quantity in Language and Thought

A one-hour slot was allocated for student sessions. In fifteen minutes, the participants were able, if desired, to present some of their research and invite other participants to collaborate in the future.

During the weekend, a *Formal Grammar 2018* conference was organised.[3]

In addition to the all aforementioned professional activities, social activities were not neglected. After the reception during the first day of the school, excursions were organised for the weekend. The first excursion led to Plovdiv, the second largest city in Bulgaria, dating from the 6th century BC. On the following day, the students could visit the monastery of Saint Ivan of Rila from the X century, located in the Rila Mountains. At the end of the first week, all participants gathered once more and enjoyed live music at the official ESSLLI party. In the second week, the participants even played a football game against the lecturers. An overview of all the activities can be found on-line.

4 Selected Courses

In this Section, several selected courses according to the author's choice will be briefly described.

4.1 Advanced Regression Methods for Linguistics

Lecturer Martijn Wieling (University of Groningen, The Netherlands) held this course in the first week. The course introduced students to advanced regression methods in *R*. The course began with a lecture on multiple regression. After that, two lectures included Gaussian and logistic mixed-effect regression, which take into account the structural variability present in the data.[4] The final two lectures of this course provided an introduction to generalised additive modeling, which is a powerful method to analyse non-linear patterns in data. This approach is especially useful when time-series data (such as EEG, eye-tracking, or articulatory data) need to be analysed.

[3] Formal Gramar Conference (on-line)

[4] For example, linguistic experiments often include participants who respond to multiple items. This structure must be brought into the model to prevent overconfident (i.e. too low) p-values.

4.2 Multiword Expressions in a Nutshell

Carlos Ramisch (Aix Marseille University, France), Agata Savary (Université François Rabelais Tours – IUT de Blois, France) and Aline Villavicencio (University of Essex, UK and Universidade Federal do Rio Grande do Sul, Brazil) held the course about one of the hottest topics in computational linguistics during the second week. The goal of this hands-on course is to provide a broad introduction to Multiword Expressions, with strong multilingual emphasis. It covered theoretical foundations, discussing properties and guidelines for their annotation, possible scenarios for their computational treatment, and techniques for idiomaticity prediction. Practical exercises provided participants with an opportunity to use different language technologies for corpus annotation and idiomaticity prediction. This course was tailored for students and researchers in computational linguistics who wish to analyse and integrate Multiword Expressions into their computational tools and linguistic studies.

4.3 Probabilistic Modeling and Bayesian Data Analysis in Experimental Semantics and Pragmatics

During the second week, Michael Franke (University of Tübingen, Germany) and Michael Henry Tessler (Stanford University, USA) posed the following questions: how do established theoretical notions lead to empirically testable predictions and what can we learn from experimental data about theoretical variables of interest? This course addresses these questions by introducing theory-driven probabilistic modeling in connection with Bayesian data analysis as a helpful set of tools to learn from observational data through the lens of a theoretical model. Lecturers introduced the basics of Bayesian data analysis and probabilistic modeling through a series of concrete case studies in natural language semantics and pragmatics.

4.4 Word Vector Space Specialisation

Ivan Vulić (University of Cambridge, UK) during the second week introduced students with the latest methods for constructing specialised vector spaces for a variety of applications in the field of natural language processing. Modern representation approaches are mainly based on the distribution hypothesis “You shall know a word by the company it keeps”, because they are based on information about the word co-occurrences in large corpora, but on other types of information, as well. Proposed approaches fall into two broad categories:

- Unsupervised methods which learn from raw textual corpora in more sophisticated ways (e.g. using context selection and attention);
- Knowledge-base driven approaches which exploit available resources to encode external information into distributional vector spaces.

The lecturer delivered a detailed survey of the proposed methods and discussed best practices for their intrinsic and application-oriented evaluation.

5 Additional Features

The organising committee of the school offers students different types of financial support each year. Grants can cover travel expenses, accommodation costs or the attendance only. When choosing grant holders, preference is given to students who actively participate in the student session, students without alternative financial support for the participation and to the outstanding and highly motivated students.

When registering on the first day, each participant received, in addition to the certificate, one paper form with empty fields for signatures. The ESSLLI organisation encourages the Universities and Educational institutions to accept ESSLLI courses as obtaining ECTS credits (at most 3 EC). After a held course, a participant can ask a lecturer for a signature. The amount of the awarding ECTS credits depends of the educational institution itself. Organisers propose that full participation in: two courses, including required reading and work, counts as 1 EC; four courses as 2 EC; and in six or more courses as 3 EC.

Next summer school will be held 5–16 August 2019 in Riga (Latvia).

COBISS Conference 2018 – 30th Anniversary

Emina Čano Tomić
emina.cano.tomic@nb.rs
National Serbian Library
Belgrade, Serbia

PAPER SUBMITTED: 12 December 2018
PAPER ACCEPTED: 14 December 2018

COBISS Conference 2018 took place from 27th to 29th December at the premises of the Institute of Information Science in Maribor (IZUM). The first conference of this kind took place in 1998, while this year's event was a celebration of the 30th anniversary of introducing the shared cataloguing system in the territory of former Yugoslavia. As a part of the development project of the Scientific and Technological Information System of Yugoslavia (SNTIJ), in 1988 the Institute of Information Science in Maribor was in charge of the task of a shared cataloguing system development which resulted in COBISS system, COBISS.SR[1] and COBISS.net[2] networks. Over the past three decades many experts from Serbia and other parts of former Yugoslavia have been working on the development and implementation of this system.

The first day of the conference was planned out for the librarians from Slovenia while the second and third day were devoted to attending guests from Albania, Bosnia and Herzegovina, Bulgaria, Montenegro, Kosovo, Macedonia and Serbia – which are all members of the COBISS.net network.

In its opening speech, Aleš Bošnjak, the Head of IZUM, announced significant changes to the system and revealed the plan of the Institute to switch to COBISS4 which is in line with the strategic decision to move the entire system to an open source platform over the next several years.

The central topic of this Conference was school library and its place within the global bibliographic information network. It was selected in order to mark the completion of the process of connecting all school libraries in Slovenia to COBISS. Participants from Slovenia emphasized all the advantages of this project for the libraries, while Dorothy Williams, professor at the Robert Gordon University in Scotland, provided a general overview,

[1] COBISS.SR (on-line)
[2] COBISS.net (on-line)

emphasizing the significance of school libraries for education, based on her extensive research in this field.

Mr. Davor Šoštarić, also from IZUM, presented the experiences of the librarians' community in Slovenia in implementation of GDPR directive regarding protection of personal data. The same directive was adopted in Serbia in 2016, and after an interim period for regulatory adjustment, it came into force in May 2018. For the time being, the number of institutions and organizations in Serbia that have successfully managed to adapt their operation to GDPR requirements is relatively small. The Directive stipulates procedures for registration, classification and coding of user data defined as personal data, whereas the participants from Serbia learned from the experience of their colleagues from Slovenia, where these provisions have been in force for some time now, which was very useful.

This year Serbia started with preparations for final implementation of the system of normative control of personal names in COBISS.SR. Persons appointed to coordinate the project are Gordana Mazić from IZUM and Milorad Vučković, the Head of the Serbian National COBISS Centre for Shared Cataloguing. IZUM representative presented an overview of the process: from decision for the initial base scope , the definition of cataloguing rules and referral sources, the definition of the language and alphabet of normative records as well as the initial integration in the first phase of development of the initial database, to testing of the software, print-out and initial connection and training of cataloguers to work in a system which includes normative control in the final phases of implementation. The task of establishing and auditing the initial normative base of personal names was assigned to the library experts working at libraries founders of VBS network in 2013. They were given the instructions and successfully provided the training to all cataloguers in normative control of personal names this year.

COBISS+ as well as the latest application for online search in library catalogues were presented at the Conference. At the beginning of 2019 this application should replace the current version of COBISS/OPAC that has been in use since 1997 with only few minor modifications. New application is simple, navigation is logical and intuitive, and interface is user friendly. It is well suited for use on mobile and touch screen devices and allows personal interface settings (My profile), logging in using multiple identities, facet filtering of hits and many other options.

The last day of the Conference, dedicated to members of COBISS.net network, began with presentation of progress reports from national COBISS centres for the last two years, followed by the presentation of the new solution for online registration of new library members and an overview of solutions used to protect the network and IZUM data.

We have seen a new generation of firewalls, solutions for early detection of weaknesses in the system and its web applications as well as zero day attack detection and prevention system.

Matjaž Zalokar (IZUM) presented the project for development of COBISS's master list of subject headings – SCG, which was used to implement the normative database of subject headings within COBISS.SI network. Work on this normative database has been going on for 18 years now and it was modelled after normative subject heading databases used by the Congress Library – SHLC and the French National Library – Rameau. After invitation of the VBS Centre, Matjaž Zalokar presented this project in the National Library of Serbia back in 2005, when it was still under development, and it was very interesting to see the project reaching its final implementation.

The Conference also included the meeting of directors of COBISS.net, as well as a number of meetings of representatives of COBISS.SR and IZUM, resulting in an agreement to have the application COBISS+ launched within COBISS.SR at the beginning of next year and to finalize implementation of normative control in March 2019.

All presentations are available at the Conference website.[3]

IZUM provided financial aid for travel and accommodation expenses for some of the participants from countries participating in COBISS.net and proved to be an excellent host, as always.

[3] The Conference website (on-line)

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).