Enrichment of Morphological Dictionary of MWUs from Library and Information Science

Aleksandra Trtovac and Jelena Andonovski

University of Belgrade, University Library "Svetozar Marković",
{aleksandra,andonovski}@unilib.bg.ac.rs
http://www.unilib.bg.ac.rs

Abstract. Terminology, as the study of terms and their use, is one of the important domains of natural language processing (NLP). It represents a very useful source of enrichment of electronic dictionaries, especially in some areas of science and technology. Terms are simple words or multi-word units (MWUs) that in specific context have a specific meaning. In this paper we will analyze terminology of library and information science, extraction of new terms from the Library and Information Science Corpus (LISC), in order to enrich of Serbian electronic dictionaries. On one hand LISC is composed of aligned texts in Serbian and English, published in journal *Infotheca*. On the other hand there are papers from library and information science which are part of the Corpus of Contemporary Serbian. We will use "Dictionary of Librarianship: English-Serbian and Serbian-English" by Ljiljana Kovačević, Vesna Injac and Dobrila Begenišić to extract new term candidates, both simple terms and MWUs. As morphological dictionaries are a necessary resource in the automatic analysis of a text we will use DELA-type dictionaries for terminology extraction and enrichment of electronic dictionary of library and information science.

Keywords: Terminology, Multi-Word Units, Electronic Dictionaries, Library and Information Science

1 Introduction

Research of natural language processing (NLP) and development of Serbian electronic dictionaries have 35 years long history at the Faculty of Mathematics, University of Belgrade. During that period, many resources for the processing of Serbian have been built by NLP group at the Faculty of Mathematics (MATF NLP), and one of most important is the system of Serbian morphological electronic dictionaries (SrpMD). SrpMD "consists of dictionaries of simple words (a sequence of alphabetical characters) and simple word forms, a dictionary of compounds (e.g. phrases and syntagms), and a dictionary consisting of finite–state transducers (FST) used for recognition of unknown words, i.e. words that are not found in other dictionaries and systems" [7].

In addition to building the electronic dictionary of simple words, many efforts were done to build SrpMD of multi-word units (MWUs). If we want to define

MWUs in an electronic dictionary, we can accept the definition given by [8] that MWUs: (a) are composed of two and more graphical words; (b) show some degree of morphological, syntactic, distributional or semantic non-compositionality; and (c) have unique and constant references. Also, MWUs are the topic of interest in many works on automatic processing of Serbian and parallel corpora [11].

The processes of information extraction, text classification, machine translation heavily rely on identification and understanding of technical terms which are semantically rich linguistic units.

Terms are simple words or MWUs that in specific context have a specific meaning. The basic facts about terms are the following: (a) new terms are constantly created with the rapid advances of science and technology; (b) most of them are nominal multi-word units (MWUs); (c) many multi-word terms contain other, previously forged, terminological MWUs [16]. Therefore terminology, as the study or the science of terms and their use, represents a very useful source of enrichment of electronic dictionaries, especially in some areas of science and technology. On the other hand, a form of electronic dictionaries is suitable for presentation and description of terminology, both simple and compound words.

In the different dictionaries, terminology is defined as "words, phrases, and symbols representing the concepts and subjects used in a specific field of research, study, or activity, for which the meaning [...] is clearly defined, sometimes in a published glossary or lexicon" [15], "the words and phrases used in a particular business, science, or profession" [1] or "the set of terms belonging to a science, a technology, or any other delineated area of thought or action." [2] Some authors claim that terminology "is not a new field of study, but only in recent decades has it been systematically developed" [3] because scientific terminology has developed during 18th and 19th century, but technical and engineering terminology had a leading role in 20th century. Terminology, as a scientific discipline, has an interdisciplinary character. Wüster [18] points up that terminology is a component part of linguistics (language units), logic and ontology (cognitive elements), information science (specialized communications), and especially in many fields of computer science. Cabré [3] stresses the importance of terminology in computer science – "the ties between linguistics and computer science have evolved gradually" and now we have a lot of computer science fields where terminology is very important – word processing, spelling checkers, electronic dictionaries, computer-assisted systems for translation, then analyzers, lemmatizers, classifiers etc. Like the other scientific fields "librarianship is one of many areas of work in which everyday exchanges include technical language, acronyms and abbreviations [13]."

1.1 Library and Information Science Terminology

Library and information science (LIS) is the scientific field that has changed its basic characteristics in the last years – from indexing, cataloguing, classification, bibliography, archiving it has moved to information technologies, information retrieval, information literacy, digitization, digital preservation, web–based

information services, management, marketing, publishing, bibliometrics and scientometrics etc. The new paradigm of LIS and its fields of practices, examples of concrete research problems, the fundamental concepts and many subareas, related disciplines, and approaches are presented in [5]. It is especially emphasized that "LIS is both a knowledge producing field and a knowledge utilizing field" and this requires a wide range of knowledge: cultural knowledge, knowledge about the different domains — philosophy and sociology of science, economic and administration, knowledge about specific information sources (databases, Internet resources), information technology, language and communication. Also, the fields and subfields of LIS changed and today it is necessary to have skills in: information retrieval, design of information systems, quality management of information services, teaching information searching, library automation, digital libraries, multimedia storage and retrieval, scientific communication etc. [5] The multidisciplinarity of this scientific field is obvious even in the curriculum at the LIS Department at the Faculty of Philology, University of Belgrade. The change of the curriculum and teaching of LIS is described in detail in [9]. The changes in LIS curriculum, especially in IT domain, are based on "dramatic changes brought to libraries and librarians by the Internet and the web, in the last two decades". As well as at universities in the EU and world wide, the LIS students at University of Belgrade have to successful adopt both theoretical knowledge and practical skills "to be advanced users of information technology and creators of digitized material". Therefore, LIS curriculum for IT contains foundation of information science — elements of mathematical logic, basics of the set theory, coding theory, statistics, data structures (theoretical knowledge) and work with standard office package, formatting programs, mark-up languages, then basics of programming, database programs and query languages (practical skills) as well as character encoding, typesetting and font design, Internet, search engines, information retrieval etc. [9] When we take a look at this curriculum, it is clear that major changes occurred in the LIS field.

In this paper we are analyzing research on terminology in LIS, in the first place MWUs. For the reasons indicated, the terminology in the field of LIS is very specific and productive because this scientific field overlaps with many other areas from archiving, documentary, history of letters and books, museology, to linguistics, computational linguistics, theory of information, programming, information retrieval, automatic text processing, etc. Also, the scientific and practical field of LIS is today inevitably supported by IT. These are the reasons for the appearance of new terms or often adoption of terms that already exist in the overlapping sciences or fields.

2 Starting Point — the Traditional LIS Dictionary

In order to enrich the dictionary, it was necessary to start with some lexical resource. For that purpose we had to use some traditional terminological dictionary with LIS terminology. Because of actuality and specificity of terminology, we decided to use a database of bilingual Serbian-English and English-Serbian

"Dictionary of Librarianship" by Ljiljana Kovačević, Vesna Injac and Dobrila Begenišić, published by the National Library of Serbia in 2004 [6]. The good side of this traditional dictionary is that in addition to the printed edition, there is an online version which can be searched in Serbian and English. This version of the Serbian-English, English-Serbian dictionary currently contains 23,400 terms, of which 12,100 are in Serbian, and 11,300 are in English. The new edition of this dictionary is supplemented by about 3,000 new entries in Serbian. Translation in English and German is in preparation. However, this dictionary has been developed independently of the projects related to the enrichment of morphological dictionary of the Serbian language.

According to [7], traditionally produced Serbian lexical resources are not suitable for automatic processing of Serbian. The advantage of the use of "Dictionary of Librarianship" is that the authors provided us with the table of terms used for the online version of this dictionary. We used this table to extract new term candidates for e-dictionaries, both simple terms and MWUs.

In the table of "Dictionary of Librarianship", the terms are divided into 50 term classes. Some of the term classes are librarianship, documentary, organization, management, information technology, cataloging, reprography, printing, electronic data processing, electronic resources, the Internet, conservation and restoration, funds, equipment, law, non-book materials, classification etc. Each term class in the table has its own code, and one term can appear in one or more classes (Table 1).

Table 1: Appearance of terms in classes

term (Serbian)	term (English)	appears in term class(es)
relativna adresa	indirect address	electronic data processing
elektronska piraterija	electronic piracy	electronic data processing; law
elektronsko izdavaštvo	electronic publishing	electronic data processing; publishing
individualni autor	individual author	librarianship; communicology; electronic data processing; cataloguing

All terms from the chosen dictionary could not be automatically processed. With manual processing it was determined that some terms must be rejected. Often certain words and phrases are quite acceptable as English terms but their translation in Serbian is not adequate either because they are not precise enough or are too descriptive. The terms beginning with KOJI (*that*) cannot be included in the Serbian dictionary as MWUs and they are rejected (Table 2).

3 Serbian Morphological Dictionary

The use of tables with classes of terms was just the first step in our research. To correctly extract new LIS terms it was necessary to use SrpMD. For terminology extraction and enrichment of an electronic dictionary with LIS terminology we used DELA-type dictionaries. These dictionaries are developed under the corpus processing system Unitex, in LADL (Laboratoire d'Automatique Documentaire et Linguistique) format. The complete work procedure in this system is described in the manual [14]. In LADL, the interest in MWUs and the production of morphological dictionaries of compounds has been vivid from the very beginning [4]. Following the same line, the system of Serbian electronic dictionaries of MWUs is being produced and it has reached the size of about 12,500 MWU lemmas [7]. One example of complementing SrpMD with MWUs is represented in [12]. The aim of our research is to increase this number.

Table 2: Rejected terms from the Serbian "Dictionary of Librarianship" beginning with "KOJI"

term (Serbian)	term (English)	class
koji upućuje na osnovno značenje	denotative	+BIINF+BILING
koji se ne može smanjiti	irreducible	+BIEOP
koji upija vodu	water-absorbing	+BIPAP+BIKONZ
koji odbija vodu	water-repellent	+BIKONZ+BIPAP
koji se može programirati	programmable	+BIEOP

The specificity of the Serbian language is the use of two alphabets (Cyrillic and Latin) and two pronunciation (Ekavian and Ijekavian). Since the texts in LISC are written in both alphabets and both dialects, we used the appropriate Serbian morphological electronic dictionaries in our research. Also, when the meaning of terms in Serbian was not clear, we used an English morphological dictionary for additional checking. For automatic extraction of MWUs we used local grammar developed under the Unitex system. In Unitex, local grammars can be represented either by a graph or by a regular expression. Our research is based both on regular expressions and graphs. The whole procedure will be described below.

4 Corpus and Verification

A corpus can be used both as a source for research and as a test area for defined models and procedures. "Text collections and corpora in digital form represent important resources for the empirical research of the Serbian language" [7].

Corpus of contemporary Serbian language (SrpKor) has been developed by the MATF NLP. The latest version of the corpus with 122 million of words is available since January 2013 [17]. Several subcorpora of Srpkor are currently used for research. The largest is an untagged corpus of contemporary Serbian language. This corpus is organised by registers and it contains texts in Serbian published in the 20th century or later. All texts represented in the corpus are devided into: newspaper texts (daily, weekly and monthly and culture supplements) dating from 1993, administrative texts, literature (both original and translated works) dating from 1920, scientific works and the other text types. The corpus is searchable by regular expressions. Another corpora available for research are: corpus with morphosyntactically tagged texts and parallel (aligned) corpora that "consist of several semantically equivalent texts that are alligned to a recognizable level that can be a paragraph, a sentence, a phrase or a word" [10].

A LISC subcorpus of SrpKor is used to verify the terms from the "Dictionary of Librarianship", to decide if term should become an entry of a Serbian morphological dictionary and to establish if a term is generaly recognised or specific to the LIS domain. The subcorpus used for our research is composed of:

1. Serbian-English aligned texts, published in the journal *Infotheca* which is issued by the Serbian Academic Library Association;
2. works in the LIS field by prof. Aleksandra Vraneš[1] and by prof. Cvetana Krstev[2].

5 Process of Extraction MWUs

In order to identify MWUs in the field of library and information science (LIS MWUs), first we used the database of "Dictionary of Librarianship" to extract tables of terms, both simple words and MWUs. Extracted terms we grouped according to the corresponding term classes and each term class was sorted in alphabetic order. Each term had a special code (tag) that indicated which class or classes the term belonged to. The first steps of LISC analysis focused on the term classes that we more related to information science and belonged to the following fields: informatics (tagged BIINF), the Internet (BIINTER), thesaurus (BITEZ), electronic resources (BIELR), search (BISSR), electronic data processing (BIEOP), linguistics (BILING).[3] It was expected that within these categories we could retrieve candidates for both the special (terminological) LIS dictionary and the general Serbian dictionary. The tags of term classes were used to

[1] Books: "Academic Libraries", "Fundamentals of Bibliography", "From Manuscripts to the Library: Glossary". Doctoral dissertation: "Serbian Bibliography in the Field of Literature, Language and Librarianship in Magazines, Newspapers, and Journals from Orfelin to 1941 : History and Theory"

[2] Materials for various courses in Computer Science and Information Science at the Department of Library and Information Science `http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1314/nastava1314.html`

[3] Prof. dr Ranka Stanković added semantic markers to term classes from Dictionary of Librarianship

extract candidates for new dictionary entries. The inspection process was done using special text editor.

The next step was a selection of MWU candidates for the dictionary. In the first step, we rejected phrases and sets of terms that could not be accepted as terms or dictionary entries. The verification of terms was done manually for each processed term class. We used Unitex system [14] in an iterative process, combining SrpMD and the incomplete LIS dictionary populated with MWU terms selected from the previously processed term classes. In each iteration we examined the produced list of unrecognized words (an Unitex error file), selected the new candidates for the LIS dictionary and repeated the process with an updated version of LISC dictionary.

During the selection of MWUs candidates, selected terms were classified into two categories. Terms used outside of the LIS domain became entries of the general dictionary, while terms used primarily by LIS were put in a special (terminological) LIS dictionary.

Since there are neither exact rules for the division on general and specific terms nor rules for the acceptance and rejection of the terms, we had to check some unclear cases in both LISC and SrpKor. More then 3000 terms were processed and more then 1000 terms were accepted as MWU candidates for DELAS lemmas (Figure 1). All inflected forms were produced for lemmas of accepted MWU terms [8].

```
sigurnosna(sigurnosni.A2:aefs1g) kopija(kopija.N600:fs1q),NC_AXN+BIeop

skener(skener.N1:ms1q) sa ravnom podlogom,NC_N6X+BIeop+BIopr

vlasnički(vlasnički.A2:adms1g)
softver(softver.N1:ms1q),NC_AXN3+BIeop+BIleg
```

Fig. 1: Examples of accepted MWU terms and their DELAS lemmas. The last example (`vlasnički softver`) should be written in one line, but we use two lines for clarity.

We will represent extraction of terms for the next four classes: BITEZ, BISSR, BILING and BIEOP. When extraction process started term classes like classification (BIKLAS), BIINTER, BIINF, indexing (BIIND) and electronic resources (BIELR) were already processed. After inspection almost all terms from the classes BITEZ and BISSR are recognized as they are already in the dictionary. For example, class BITEZ had 17 simple terms and 87 MWUs at the beginning. The inspection process established that 15 simple terms and 80 MWUs were already in the dictionary. From the remaining 7 MWUs 4 were rejected because they were too descriptive and 3 were classified for the general dictionary: terms "ciljni jezik" (target language), "istorijska napomena" (history note) and "složeni termin" (compound term) were classified for the general dictionary

while terms "nepotpuna ekvivalentnost" (inexact equivalence), "privremena predmetna odrednica" (provisional term), "privremeni deskriptor" (provisional term) and "uputno-informacioni sistem tezaurusa" (lead-in vocabulary) were rejected. As the number of terms was not large, the process of classification was done easily and quickly.

But the situation with term classes BILING and BIEOP was a bit different. These classes were bigger and more complex, especially BIEOP, and we needed more time to process it. The inspection process was done easily and one third of all terms was recognized as already included in the dictionary. But for the examination process whether terms were for general or special (terminological) LIS dictionary we needed more time than for the previous two classes. For example, the class BILING had 135 simple terms and 246 MWUs. The inspection established that 53 simple terms and 120 MWUs were already in the dictionary. Of the remaining 82 simple terms, 24 were selected as the candidates for new dictionary entries, while 58 simple terms were rejected. It was determined that 82 of 126 MWUs were candidates to be included in the dictionary, while 44 MWUs were rejected. The examples of terms included in the special (terminological) LIS dictionary are "jezik katalogizacije" (cataloguing language), "jezik uporednog stvarnog naslova" (language of parallel title proper), while the new entries of the general dictionary are the terms "krsno ime" (Christian name), "maternji jezik" (native language), "pozajmljena reč" (loan–word) (Table 3).

Table 3: Accepted terms for the dictionary (general or special) from the BIling class

term (Serbian)	term (English)	class
jezik katalogizacije	cataloguing language	+BIBIBL+BIKAT+BILING
jezik uporednog stvarnog naslova	language of parallel title proper	+BIKAT+BILING
krsno ime	Christian name	+BIGEN+BILING
maternji jezik	native language	+BILING
pozajmljena reč	loan-word	+EK+BILING

The largest class of all was the electronic data processing (BIEOP). At the beginning this class had 615 simple terms and 1599 MWUs. The inspection process showed that 117 simple terms and 257 MWUs were already in the dictionary. The rest of 498 simple terms and 1342 MWUs were examined for their appearance in LISC, as well as their concordances of SrpKor. Total of 57 simple terms and 1084 MWUs were selected as the new entries of SrpMD (Table 4), 641 MWUs as the entries of general dictionary and 443 MWUs as the entries of the special (terminological) LIS dictionary.

Table 4: Candidates for the dictionary from the BIeop class

term (Serbian)	term (English)	class
automatsko prevođenje	computer aided translation (CAT) / automatic translation	+BIeop+BIling
digitalna biblioteka	digital library	+BIbibl+BIeop+BIorg
ekran osetljiv na dodir	touch screen / touch-sensitive screen	+Ek+BIeop
elektronsko izdavaštvo	e-publishing / electronic publishing	+BIeop+BIpubl
geografski informacioni sistem	geographic information system (GIS)	+BIeop+BIkart
jezik za označavanje hiperteksta	Hypertext Markup Language (HTML)	+BIeop+BIling
mašinski čitljiv katalog	machine-readable catalogue	+BIeop+Bikat
obrada prirodnog jezika	natural language processing	+BIeop+BIling
personalni računar	personal computer - PC	+BIeop+BIopr
poštansko sanduče	mailbox	+BIeop+BIkom

Examples of terms selected for the general dictionary terms are "administrator sistema" (system administrator), "bežični LAN" (wireless LAN), "digitalna zbirka" (digital collection), "elektronska obrada podataka" (electronic data processing, EDP), "nalepnica sa bar kodom" (barcoding label). Some of the terms included in the special (terminological) LIS dictionary are "automatizovana katalogizacija" (computerized cataloguing), "blok uputnice "vidi"" (see reference tracing block), "eksport podataka" (data export), "normativni zapis" (authority entry record), "računarska lingvistika" (computational linguistics) (see (Table 5)). The examples of rejected terms are given in (Table 6).

Examination process was based on a text search with Unitex finite state transducers (FSTs) and analysis of produced concordances. The searched texts were previously processed with a combination of SrpMD and current version of LISC dictionary. One way of specifying the Unitex FST for the examination process is to use a Unitex regular expression like the following[4]:

```
<N+BIelr> + <N+BIind> + <N+BIinf> + <N+BIinter> + <N+BIklas>
+ <N+BIling> + <N+BIssr> + <N+BItez> + <N+BIeop>
```

(LIS-RE)

With regular expression (LIS-RE) we are looking in a text for the occurrences of nouns with one of the special markers (meaning they represent a term from

[4] The regular expression (LIS-RE) should be written in one line. Here we use two lines for clarity.

Table 5: Accepted terms for general or special (terminological) LIS dictionary from the BIeop class

term (Serbian)	term (English)	class
administrator sistema	system administrator	+BIeop+BIpers
bežični LAN	wireless LAN	+BIeop+BIkom
digitalna zbirka	digital collection	+BIeop+BIfond
elektronska obrada podataka	electronic data processing (EDP)	+BIeop
nalepnica sa bar kodom	barcoding label	+Ek+BIeop
automatizovana katalogizacija	computerized cataloguing	+BIeop+BIkat
blok uputnice "vidi"	see reference tracing block	+BIeop+BIkat
eksport podataka	data export	+BIeop
normativni zapis	authority entry record	+BIeop+BIkat
računarska lingvistika	computational linguistics	+BIeop+BIling

Table 6: Rejected terms from the BIeop class

term (Serbian)	term (English)	class
horizontalna orijentacija papira pri štampanju	landscape mode	+BIeop+BIprint
isecanje i prenošenje	cut and paste	+Ek+BIeop
isključivo tekstualna datoteka	text-only file	+BIeop
metaoznaka polja u hipertekstu	metatag	+BIeop
obezbeđenje pristupa regionalnoj mreži	wide area providing (WAP)	+Ek+BIeop+BIkom

LIS domain's class). The results of this examination process are represented in the Table 7. The appearance of MWUs in LISC is represented by concordances of SrpKor in the Figures 2a and 2b.

As for texts published in the journal *Infotheca*, the examination process was done in the same way as with the LISC and the appearance of MWUs is represented by concordances in the Figure 2c. Journal *Infotheca* is bilingual and analysed texts are Serbian-English aligned texts, but since our research concerns with the SrpMD, we focused only on the version of the texts in Serbian.

The second way of specifying the Unitex FSTs is based on Unitex graphs which represent another form of regular expressions. Graphs can be used to construct local grammars that are powerful tool to represent majority of linguistic phenomena [14]. For example, we have tested a graph that looks for the compound constructions of the form "possessive adjective-noun" in our LISC corpus

Table 7: Examples of terms from the BIeop class separated weather for general or special (terminological) LIS dictionary

texts from LISC	recognized LIS terms	class
Serbian Bibliography in the Field of Literature, Language and Librarianship in Magazines, Newspapers and Journals from Orfelin to 1941:		
History and Theory	62	76
From Manuscripts to the Library: Glossary	185	213
Fundamentals of Bibliography	84	97
Academic Libraries	169	175
Baze_Uvod	16	17
Internet	81	90
Računar	32	72
Softver	17	40

and have analysed concordances in order to detect terms not yet present in the SrpMD. Materials for courses in computer science and library informatics gave little or no results (Figure 3a). On the other hand, works by prof. Aleksandra Vraneš ("From manuscripts to the library: glossary" and "Fundamental of Bibliography") proved to be a significant source for retrieving new terms, since they produced considerable number of useful concordances (Figures 3b, 3c).

6 Conclusion

Classified terms from BItez, BIssr, Biling and BIeop classes are now in DELAC dictionary. Dictionary is enriched with more then 1000 MWUs from these term classes. Including previously processed term classes DELAC dictionary now has 2318 MWUs. New dictionaries were tested on texts from LIS domain which are part of SrpKor and some results are represented in this paper. Although there is a considerable work to be done concerning the unprocessed classes of words, we believe that a large part of the identified MWUs can become an integral part of the general Serbian dictionary in everyday use. Also, morphological dictionary of MWUs produced from LIS terminology will be a good source for further research in this area. There are more term classes related to different fields with potential candidates for new entries in general Serbian dictionary and special (terminological) LIS dictionary. Analysis of these term classes will be a part of our further research in this area.

prostora uzrokovale su pojavu abrevijacija, skraćenica uz
njegovoj distribuciji staraju agregator servisi. Stoga se
rema po-stojećim pravilima za arhivalije.S Kompletnost opi
aj knjižni, rukopisni i drugi arhivski fond ili dati deskr
stupno je 12000 časopisa, kao baza punog teksta i baza saž

(a) "From manuscripts to the library: glossary", prof. Aleksandra Vraneš

yu je oznaka države i to je domen najvišeg nivoa. Poseb
iz padajućeg menija izaberemo format datoteke, recimo "Adobe
gram traži od lokalnog servera imena domena da mu da IP adres
net je danas široko raširena informaciona infrastruktura.
etraživati menije; - koristiti klijentski program Gopher koji

(b) The materials for courses in computer science and library informatics at the Department of Library and Information Science, prof. Cvetana Krstev

juma META-RAZMENA.</seg> <seg>Ciljna zajednica korisnika pra
zitetu u Plovdivu (na primer, elektronski rečnici), Novom bu
anju funkcionalnosti umrežene informacione tehnologije.</seg
dinstveno digitalno tržište i informacioni prostor.</seg></p
tečeno je značajno iskustvo u istraživačkom radu i postignut

(c) The *Infotheca* article: Maciej Ogrodniczuk et al. Jezički resursi Centralne i Južne Evrope u okviru platforme META-RAZMENA. *Infoteka*, 13(1):3–28, 2012.

Fig. 2: Examples of concordances produced by search in LISC and *Infotheca* using regular expression (LIS-RE).

References

1. Macmillan Dictionary. Terminology.
2. ODLT – The Online Dictionary of Language Terminology. Terminology.
3. Maria Teresa Cabré. *Terminology*. John Benjamins Publishing Company, Philadelphia, PA, USA, 1999.
4. Maurice Gross. Lexicon–Grammar. The Representation of Compound Words. In *Proceedings of Coling 1986*, pages 1–6, Bonn, 1986.
5. Birger Hjørland. Library and Information Science: Practice, Theory, and Philosophical Basis. *Information Processing & Management*, 36(3):501–531, 2000.
6. Ljiljana Kovačević, Vesna Injac, and Dobrila Begenišić. *Dictionary of Librarianship: Serbian-English, English-Serbian*. National Library of Serbia, Belgrade, 2004.
7. Cvetana Krstev. *Processing of Serbian – Automata, Text and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade, 2008.
8. Cvetana Krstev, Ranka Stanković, Ivan Obradović, Duško Vitas, and Miloš Utvić. Automatic Construction of a Morphological Dictionary of Multi-Word Units. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Proceedings of the 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010*, number 6233 in Lecture Notes in Computer Science, pages 226–237, Heidelberg, 2010. Springer.

a DBTG (Data Base Task Group) američke organizacije CODASYL
parentna. To znači da promene fizičke organizacije podataka
i se zasnivao na ovom modelu. Hijerarhijske veze između raz
jekta njegovim sve podrobnijim informatičkim aproksimacijama.
nja za izradu aplikacija kao i korisničke sumeđe koja sloboda

(a) The materials for courses in computer science and library informatics at the Department of Library and Information Science, prof. Cvetana Krstev

rna hartija, pravljena od afričke biljke halfe, koja se
nsku crkvu. Otvaranje nove Aleksandrijske biblioteke inic
u kojoj se upotrebljavaju anilinske boje, prozirne i nem
e, evaluacije i korišćenja arhivske građe, kao i istorije
tikovana, te ih nacionalne bibliografske agencije nisu u

(b) "From manuscripts to the library: glossary", prof. Aleksandra Vraneš

(Nj. Pool) u bibliografiji Alfabetski indeks tema, obrađ
edne, ili više biblioteka (centralni katalog), ako su u
jevačkom Liceumu, uredniku ekonomskog dodatka Srpskih no
sačuvan u legatu profesora Filološkog fakulteta Đorđa Ži
ivanja građe, alfabetski i hronološki pristup, doduše fo

(c) "Fundamental of Bibliography", prof. Aleksandra Vraneš

Fig. 3: Examples of concordances produced by testing the Unitex graph that looks for the compound constructions of the form "possessive adjective-noun" in LISC.

9. Cvetana Krstev and Aleksandra Trtovac. Teaching Multimedia Documents to LIS Students. *Journal of Academic Librarianship*, 40(2):152–162, 2014.
10. Cvetana Krstev and Duško Vitas. An Aligned English-Serbian Corpus. In N. Tomović and J. Vujić, editors, *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I, Belgrade, 4-6 December 2009*, pages 495–508, Belgrade, 2011. Faculty of Philology.
11. Cvetana Krstev, Duško Vitas, and Aleksandra Trtovac. Orwell's 1984 – the Case of Serbian Revisited. In Z. Vetulani, editor, *Proceedings of 5th Language & Technology Conference, November 25-27, 2011, Poznań, Poland*, pages 570–574, Poznań, 2011. Fundacja Uniwersytetu im. A. Mickiewicza.
12. Cvetana Krstev, Duško Vitas, and Aleksandra Trtovac. Orwell's 1984 – From Simple to Multi-word Units. In *Human Language Technology Challenges for Computer Science and Linguistics, LNCS*, pages 276–287. Springer International Publishing, 2014.
13. Mary Mortimer. *Library Speak : A Glossary of Terms in Librarianship and Information Management*. Total Recall Publications, Friendswood, TX, USA, 2007.
14. Sébastien Paumier. *Unitex 3.0 User Manual*, 2011.
15. Joan M. Reitz. Online Dictionary for Library and Information Science. Terminology.

16. Agata Savary, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek, and Filip Makowiecki. SEJFEK – a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III) COLING 2012, Mumbai, December*, pages 195–214, 2012.
17. Miloš Utvić. *Izgradnja referentnog korpusa savremenog srpskog jezika: doktorska disertacija*. PhD thesis, Univerzitet u Beogradu, Filološki fakultet, Beograd, 2013.
18. Eugen Wüster. L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In H. Rondeau, G. & Felber, editor, *Textes choisis de terminologie*, volume Vol. 1: Fondements théoriques de la terminologie, pages 55–113, Ville de Québec, 1981. Université Laval - GIRSTERM.