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

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

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

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 (entire document), <body> (header), <p> (paragraphs) and <seg> (sentences).

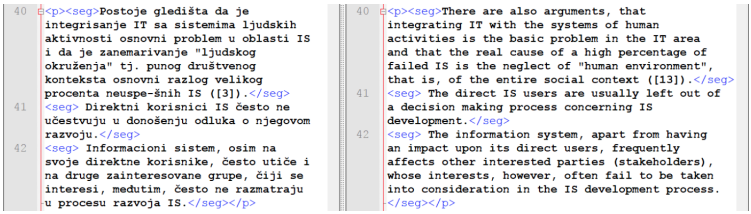

Figure 1. An example of a prepared parallel English-language text in the TEI/XML format

.

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).

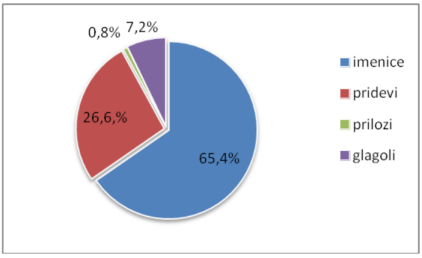

Figure 3. Word class distribution in the set of 500 extracted key lexical units

.

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 ...	n78 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 agencie.	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 ...	n160 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.

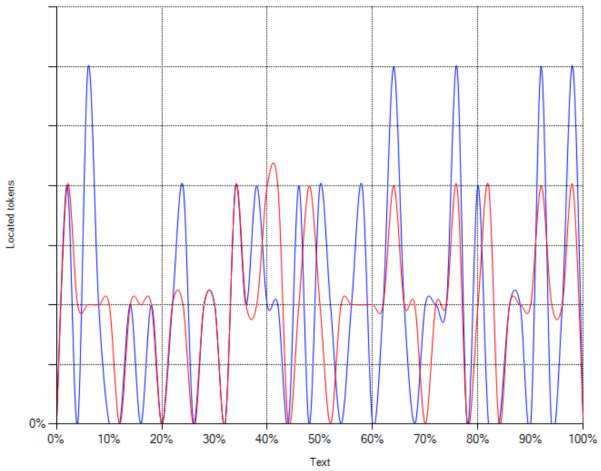

Figure 4. Distribution of the Serbian *менаџмент* и the English *management* in the corpus

.

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