Rule-based Automatic Multi-Word Term Extraction and Lemmatization

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, Aleksandra Trtovac

Rule-based Automatic Multi-Word Term Extraction and Lemmatization

Ranka Stanković[1], Cvetana Krstev[2], Ivan Obradović[1], Biljana Lazić[1], Aleksandra Trtovac[3]
[1]University of Belgrade, Faculty of Mining and Geology
[2] University of Belgrade, Faculty of Philology
[3] University Library "Svetozar Marković", Belgrade
E-mail: ranka.stankovic@rgf.bg.ac.rs, cvetana@poincare.matf.bg.ac.rs, ivan.obradovic@rgf.bg.ac.rs, biljana.lazic@rgf.bg.ac.rs, aleksandra@unilib.bg.ac.rs

Abstract

In this paper we present a rule-based method for multi-word term extraction that relies on extensive lexical resources in the form of electronic dictionaries and finite-state transducers for modelling various syntactic structures of multi-word terms. The same technology is used for lemmatization of extracted multi-word terms, which is unavoidable for highly inflected languages in order to pass extracted data to evaluators and subsequently to terminological e-dictionaries and databases. The approach is illustrated on a corpus of Serbian texts from the mining domain containing more than 600,000 simple word forms. Extracted and lemmatized multi-word terms are filtered in order to reject falsely offered lemmas and then ranked by introducing measures that combine linguistic and statistical information (C-Value, T-Score, LLR, and Keyness). Mean average precision for retrieval of MWU forms ranges from 0.789 to 0.804, while mean average precision of lemma production ranges from 0.956 to 0.960. The evaluation showed that 94% of distinct multi-word forms were evaluated as proper multi-word units, and among them 97% were associated with correct lemmas.

Keywords: term extraction, terminology, multi-word units, lemmatization, finite-state transducers

1. Motivation

Various approaches have been proposed for multi-word term (MWT) extraction as this problem has been gaining in importance in the field of Natural Language Processing. Initially, MWT extraction from domain texts has been tackled mainly using the statistical approach based on different statistical measures, following the seminal work of Kenneth Church and Patrick Hanks (1990; 1991) and Frank Smadja (1993). A language independent statistical corpus based term extraction algorithm used on English and Chinese corpora is described in (Pantel&Lin, 2001), while Chen and his associates present a MWT extraction system based on co-related text-segments within a set of documents (Chen et al., 2006). Statistical measures of co-occurrence (MI^3 – mutual information) were used for finding MWT candidates in Croatian texts (Tadić&Šojat, 2003).

Although the statistical approach has been steadily pursued by a number of researchers, development of lexical resources and local grammars has given impetus to an alternative approach, namely multi-word extraction based on linguistic rules. Recently, a rule-based approach for the extraction of terms was successfully applied on an Arabic scientific and technical corpus, using a cascade of transducers (Ammar et al., 2015). Another example of this approach, SEJFEK, consisting of a grammatical lexicon of about 11,000 Polish MWTs from the economical domain, where inflectional and syntactic variations are described via graph-based rules, is described in (Savary et al., 2012).

However, the two approaches are more and more often combined in a hybrid approach. An approach to extracting MWTs from Arabic specialized corpora that uses linguistic rules to parse documents and retrieve candidate terms and statistical measures to deal with ambiguities and rank candidate terms is given in (Bounhas&Slimani, 2009). Several hybrid methods for extraction of MWT candidates that use both syntactic patterns and statistical measures, mainly for filtering, are described in (Koeva, 2007) for Bulgarian, in (Vintar, 2010) for Slovene and in (Broda et al., 2008) for Polish.

MWT candidates are extracted from texts in various inflected forms. Nevertheless, lemmatization of extracted MWT candidates, that is, their linking to one normalized or head-word form, has attracted less interest, no doubt because it is of little importance for English. For instance, the goal of Schone and Jurafsky (2000) was "to identify an automatic, knowledge-free algorithm that finds all and only those collocations where it *is* necessary to supply a definition." However, in order to accomplish this complex task authors made little if any effort to normalize extracted MWT candidates: "Prior to applying the algorithms, we lemmatize using a weakly-informed tokenizer that knows only that white space and punctuation separate words."

However, for highly-inflected languages, such as Serbian and other Slavic languages, this task can hardly be avoided as each nominal MWT can have many inflected forms (from five to ten or even more) and many of these forms (but usually not all) can in general be extracted from a corpus. If some statistical approach is used for term extraction, then at least a simple form lemmatization should be performed in order to obtain some kind of a normalized form to which all inflected forms should map. This normalized form is of no use, however, if human evaluation of results is to follow and if the approved MWT is to be entered into some kind of a dictionary or a terminological data-base. In this case we need a lemmatized MWT, that is, a MWT in the form of a dictionary head-word.

The problem of lemmatization of special kind of MWUs,

person names, was tackled for Polish (Piskorski et al., 2007). The authors used several statistical approaches that outperformed the approach relying on heuristics and linguistic knowledge, presumably because linguistic resources and tools they used were underdeveloped. In (Małyszko et al., 2015) authors lemmatize multiword entity names (organization names and similar named entities found in a corpus of legislative acts) by using rules generated on the basis of corpora analysis.

For tackling the problem of MWT extraction and lemmatization from Serbian texts we have chosen a rule-based approach, which relies on a system of language resources such as morphological e-dictionaries and grammars developed within the University of Belgrade Human Language Technology Group (Vitas et al., 2012). For our approach, production of lemmas for various forms of MWTs extracted from a corpus is necessary for two main reasons. Firstly, the evaluators need to be supplied with correct lemmas in order to be able to accomplish their task successfully. Secondly, lemmas are necessary for incorporating the MWTs in morphological dictionaries in compliance with the form these dictionaries require. This is essential as the set of forms found in the corpus is rarely comprehensive, and thus all potential forms of the term can be generated only from a lemma.

Our paper is organized as follows. In Section 2 we present the methodology and design of our system, in Section 3 its architecture is outlined, while in Section 4 we present the evaluation procedure and its results on a domain corpus. Finally, in Section 5 we give some concluding remarks and present plans for system improvement.

2. Methodology and Design

Entries in the Serbian e-dictionary of general multi-word units (MWUs) are classified according to their syntactic structure and inflectional and other properties (omission of a constituent, reverse order, exchangeability of constituent separators, etc.). Class names correspond to FSTs used for inflection of MWUs belonging to that class. For example, MWUs composed of an adjective (A) followed by a noun (N), which agree in gender, number, case and animateness, belong to the AXN class. X stands for a component that does not inflect when the MWU inflects or a separator, usually a space or a hyphen.

Nominal MWUs in Serbian belong to one of several tens of different general classes, but 14 of these classes account for more than 98% of all nominal MWUs. Four of them contain two component MWUs, five contain 3-component MWUs and four contain 4-component MWUs. As the thirteen classes cover the large majority of MWUs, lexical rules and the corresponding finite state transducers (FSTs) have been developed for the extraction of MWTs belonging to these classes, with the assumption that structures used most frequently for general MWUs would be the most frequent for terminological MWUs as well. Details on these classes are given in (Krstev et al., 2015).

The FST graph for extraction of NXN type MWUs (a noun followed by a noun that agrees with it in number and case, where the separator can be a hyphen) is depicted in Figure 1 (top). It should be noted that this graph, as all other extraction graphs, works locally, that is, it does not look at the broader context. Two subgraphs, NNp and NNs, recognize possible singular and plural forms of such MWUs, respectively. Dictionary variables $n1.LEMMA$ and $n2.LEMMA$ at FST output perform normalization, that is, simple word lemmatization by retrieving lemmas for the recognized word forms $n1$ and $n2$.

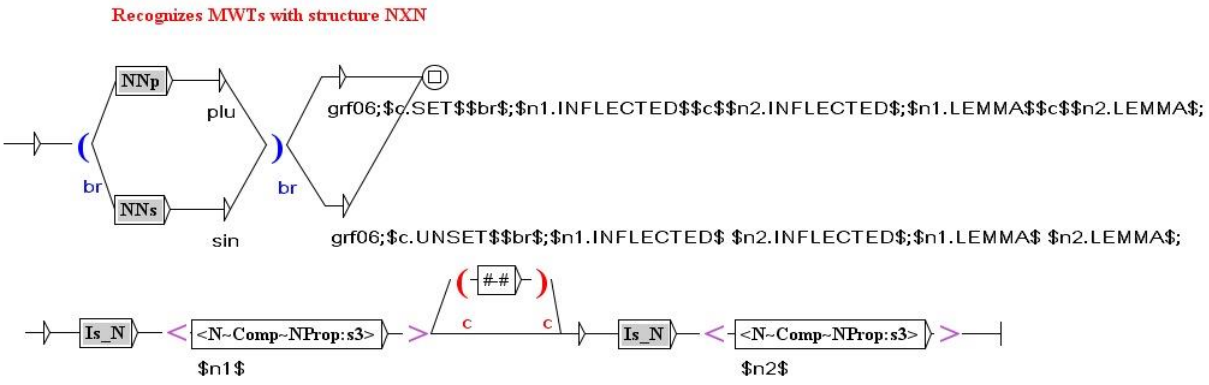


Figure 1. The FST for extraction of MWUs with the structure NXN

The output variable br produces the potential grammatical number of the recognized construction. For instance, the MWT *mašina taložnica* 'jig' has 7 different inflected forms: *mašina taložnica*, *mašine taložnice*, *mašini taložnici*, *mašinu taložnicu*, *mašino taložnice*, *mašinom taložnicom*, *mašinama taložnicama*. However, most of these forms can represent several sets of grammatical values, e.g. *mašine taložnice* can be a singular form in the genitive, or a plural form in the nominative, accusative of vocative, while *mašinu taložnicu* can be only in the singular (accusative case) and *mašinama taložnicama* only in the plural (dative, instrumental or locative). This information on potential grammatical number is recorded in the output variable

br and subsequently used for MWU lemmatization as for each MWU lemma it is necessary to determine its grammatical number (e.g. the lemma *klešta papagajke* 'parrot nose pliers' from the class NXN is in the plural). The name of the FST (grf06) that recognized the MWU is added to the output as it is used in subsequent phases to identify the MWU's syntactic structure. Figure 1 (bottom) shows one path from the NNs subgraph for the agreement between two nouns in gender and number. The subgraph Is_N within this path rejects nouns that are homographous with other PoS word forms in order to avoid false recognitions.

Given the high level of homography of word forms in Serbian it is possible that two or more graphs recognize the same word sequence where only one of them is correct. In the case of such ambiguous recognition precedence is always given to the more probable case according to the predefined order of precedence of graphs and/or frequency of candidate lemmas. For instance, in the case of 2-component MWUs the order of precedence of graphs is: AXN, 2XN (a noun preceded by a word that does not inflect in the MWU), N2X (a noun followed by a word that does not inflect in the MWU), NXN. Thus, two MWU forms *mašine taložnice* and *korita trake* 'belt troughs' would be both extracted by graph 03 (structure N2X) and graph 06 (structure NXN). Precedence would be given to the output of graph 03 because the structure N2X is much more frequent than NXN. For these two examples, it would be correct for the second MWU but not for the first. It should be noted, however, that precedence determines only the rank of a MWU lemma in a list of lemmas prepared for the evaluation, without deleting any candidate lemmas from it.

In general, the longest match for the MWU is looked for. For example, if a text sequence matches the AXAXN pattern (a noun preceded by two adjectives that agree with it in gender, number, case and animateness), then a lower rank will be assigned to the subsumed match AXN in the final phase of disambiguation. For example, the MWT *geološki informacioni sistem* (AXAXN) 'geological information system' would be given precedence over *informacioni sistem* (AXN) 'information system'. However, in this case, both would be accepted as MWT.

Extraction graphs perform simple word lemmatization the result of which need not be a correct MWU lemma. As evaluation of retrieved MWUs can only be based on correct MWU lemmas, another suit of FSTs produces candidates for correct lemmas. Our results show that correction is needed for approximately half of all candidates. The most important cases when correction of simple word lemmatization has to be done are: (a) a head-noun of a MWT is not masculine, so the adjective simple word lemma (always given in the masculine gender) does not agree with it; (b) a MWT lemma should be in the plural, although singular forms exist for simple word constituents. The first case occurs with the MWT *električna.f energija.f* 'electric energy' whose simple word lemma is **električni.m energija.f*. The second case occurs with the MWT *minerski.p radovi.p* 'blasting works' whose simple word lemma is **minerski.s rad.s*.

These correct lemma candidates, provided with information about their syntactic structures, are used to fully automatically produce entries in morphological e-dictionaries according to a strategy for producing MWU lemmas, and subsequently all their inflectional forms (for more details see (Krstev et al. 2013)).

Some extracted MWU forms might have several candidate lemmas assigned to them (due to recognition by different graphs and/or due to homography of simple words). In such cases some heuristics is used to eliminate some false suggested lemmas and to appropriately rank the remaining ones based on the number of different word forms retrieved. This problem and its solution will be explained on one particularly complex example.

The MWT *obloga trake* 'belt coating' has the structure N2X (a noun followed by a noun in the genitive), and its components are nouns *obloga* 'coating' and *traka* 'belt'. However, some of its inflected forms are homographous with other nouns, namely *oblog* 'stupe' and *trak* 'tentacle', yielding various interpretations for various inflected forms, as presented in Table 1. Data in Table 1 show that only if forms *oblogo trake* and *oblogama trake* are extracted from a text a correct lemma can be associated with certainty. But if *obloge trake* is extracted, then two different structures – N2X and NXN – can be associated with various interpretations for both constituents. However, if some other form is extracted besides it, e.g. *oblozi trake*, then some of these false interpretations would be removed – *oblog* for the structure N2X, and *obloga trak* and *obloga traka* for the structure NXN.

MWT inflected form	Grammatical values	Struc.	First constituent	Second constituents
obloga trake	s_nom+ p_gen	**N2X**	**obloga_s_nom**	**traka_s/p_gen**
			obloga_p_gen	traka_s/p_gen
			oblog_s/p_gen	traka_s/p_gen
		NXN	oblog_s_gen	traka_s_gen
obloge trake	s_gen+ p_nom/ acc/voc	**N2X**	**obloga_s_gen+ p_nom/ acc/voc**	**traka_s/p_gen**
			oblog_p_acc	traka_s/p_gen
		NXN	oblog_p_acc	trak_p_acc
			oblog_p_acc	traka_p_acc
			obloga_p_acc	trak_p_acc
			obloga_s_gen+ p_nom/acc	traka_s_gen+ p_nom/acc
oblozi trake	s_dat/loc	**N2X**	**obloga_s_dat/ loc**	**traka_s/p_gen**
		NXN	oblog_p_nom/voc	traka_p_nom/ voc
oblogu trake	s_acc	**N2X**	**obloga_s_acc**	**traka_s/p_gen**
			oblog_s_dat/ loc	traka_s/p_gen
oblogotrake	s_voc	**N2X**	**obloga_s_voc**	**traka_s/p_gen**
oblogom trake	s_ins	**N2X**	**obloga_s_ins**	**traka_s/p_gen**
			oblog_s_ins	traka_s/p_gen
oblogama trake	p_dat/ ins/loc	**N2X**	**obloga_p_dat/ ins/loc**	**traka_s/p_gen**

Table 1. A MWT *obloga trake*, its possible inflected forms, and their interpretations by extraction graphs and e-dictionaries. In bold are highlighted correct interpretations.

A similar input-driven approach is used to determine

possible grammatical number of a MWT, and consequently that of a lemma. For instance, if *obloge trake* and *oblogom trake* are extracted from a text, then we know that the MWT has singular forms (due to the second MWT) and the lemma has to be in the singular. But, if only the form *oblogama trake* is extracted we know that the MWT has plural forms and we presume that the lemma is also in plural; however, a lemma in singular remains as a possibility.

From our corpus we extracted only one form for this MWT: *obloga trake*, which did not enable elimination of any false recognitions. In this case we could only rank them – options N2X first, and options NXN later. Such highly ambiguous cases are, however, not very frequent.

The whole process, from extraction to e-dictionary production will be illustrated with another example of MWT extracted from our evaluation corpus – *električna energija* 'electric energy'. The graph that retrieves syntactic constructions AXN (grf01) extracted four different forms – three of them definitely singular and one that can be both singular and plural. The temporary lemma obtained by simple word lemmatization is not correct in either case (singular and plural), thus correction is needed (presented in column 'Lemma' of Table 2). In the filtering phase, since forms were retrieved that are undoubtedly singular, the singular form lemma is retained. This form, together with information about its structure is enough for automatic production of the e-dictionary lemma (DELAC), while information provided with the lemma enables subsequent production of all its inflected forms (DELACF).

Graph	Num	Recognized form	Frequency	Temporary lemma	Lemma
grf01	plu	električne energije	85	električni energija	električne energije
grf01	sin	električna energija	10	električni energija	**električna energija**
		električne energije	85		
		električnom energijom	8		
		električnu energiju	5		

DELAC	električna(električni.A2:aefs1g) energija(energija.N600:fs1q),NC_AXN
DELACF	električnoj energiji,električna energija.N:fs7q
	električne energije,električna energija.N:fp1q
	električnih energija,električna energija.N:fp2q
	električnim energijama,električna energija.N:fp3q

Table 2. The MWT *električna energija*, its recognized forms and lemma

The set of lemmas produced in the aforementioned filtering procedure is further processed by introducing measures that combine linguistic and statistical information. Namely, for each lemma, besides frequency, the basic measures (C-Value T-Score, LLR, and Keyness) (Frantzi, 2000; Dunning, 1993; Kilgarriff, 2014) and pondered measures that combine them are calculated. Based on a chosen measure and the corresponding threshold, the set of lemmas for evaluation is generated. Results of frequency, C-Value and T-Score measures illustrate the MWU lemma rank within the domain corpus, whereas the remaining two measures compare term frequency in the domain corpus and the general language corpus, thus illustrating how specific the MWU is for the selected domain. As the general corpus we used a 22 million words excerpt from the Corpus of Contemporary Serbian (SrpKor – http://www.korpus.matf.bg.ac.rs). The computed basic measures and their rank for the example term *električna energija* as well as some other terms are given in Table 3.

				Measures					Rank				
Graph	Num	Lemma	Eliminates lemma	Freq	CValue	TScore	LLR	Keyness	Freq	CValue	TScore	LLR	Keyness
grf01	sin	mineralna sirovina	mineralne sirovine	736	736	27129	5144	707	1	1	1	1	1
grf01	sin	površinski kop	površinski kopovi	305	230	17464	1925	161	2	4	2	3	17
grf01	sin	toplotni tok		258	249	16062	1803	335	4	3	4	4	3
grf01	sin	toplotna provodljivost	toplotne provodljivosti	236	222	15362	1649	312	5	5	5	5	4
grf03	sin	kvalitet uglja	kvaliteti uglja	297	289	17234	2076	375	3	2	3	2	2
grf03	sin	nivo buke	nivo buka;nivoi buke	118	118	10863	825	171	15	18	15	13	11
grf01	sin	površinska povreda	površinske povrede	116	104	10770	811	169	17	27	17	15	13
grf09a	sin	površinska povreda u predelu	površinske povrede u predelu	24	48	4899	168	39	272	90	272	262	223
grf05	sin	površinska povreda potkolenice	površinske povrede potkolenic	8	13	2828	56	14	1660	867	1660	1544	1521
grf01	sin	električna energija	električne energije	121	121	11000	128	4	13	15	13	383	1739
grf01	sin	električna provodljivost	električne provodljivosti	46	46	6782	321	72	97	94	97	87	80
grf03	sin	vibracija šake	vibracije šake	54	34	7348	377	83	74	171	74	66	62
grf04b	sin	vibracija šake ruke		54	63	7348	377	83	75	56	75	67	63
grf08b	sin	sindrom vibracija šake ruke	sindromi vibracija šake ruke	16	32	4000	112	26	539	190	539	514	496
grf04b	sin	sindrom vibracija šake	sindromi vibracija šake	16	0	4000	112	26	551	1690	550	496	489
grf04b	sin	merenje vibracija šake	merenja vibracija šake	15	0	3873	105	25	577	1691	575	557	533
grf08b	sin	merenje vibracija šake ruke	merenja vibracija šake ruke	15	30	3873	105	25	584	204	584	547	529

Table 3. Ranking of retrieved MWU forms and lemmas

Površinska povreda 'surface injury' was ranked lower by CValue (27) than by Freq (17) because it is part of two other MWUs with three components. One of them, *površinska povreda potkolenice* 'surface lower leg injury' was ranked higher by CValue (867) than by Freq (1660), and as this term is more common in the mining corpus than in SrpKor, it was ranked higher by LLR (1544) and Keyness (1521) than by Freq (1660). A similar example is *vibracija šake* 'hand vibration' and MWUs that contain it. On the other hand, *električna energija* 'electrical energy' is more frequently mentioned in SrpKor (a lot of news articles), and thus in terms of domain specificity it is being pushed to the bottom of the list (Keyness=1739). *Nivo buke* 'noise level' is better ranked by LLR (13) and Keyness (11) compared to Freq (15) due to a multitude of books in the field of occupational safety in mines, and the fact that the term 'noise level' is characteristic for mining.

3. System Architecture

The automatic procedure for MWT extraction and lemmatization is implemented using Unitex, a corpus processing system (http://www-igm.univ-mlv.fr/~unitex), and LeXimir, a multipurpose software tool for language resources management, developed within the University of Belgrade HLT group (Stanković et al. 2011). The whole process is automated, and takes place with very little human intervention, starting from the tokenization and lexical analysis of a raw text up to production of dictionary entries.

The system relies Unitex routines for text analysis and FST application, while one of the many functionalities of LeXimir is used to produce dictionary lemmas in various standard formats, such as LMF or TBX. All results and corresponding metadata are stored in a SQL Server database.

The architecture of the software solution depicted in Figure 2 is based on web services, thus enabling other applications to use some of them, such as indexing or document information retrieval, for term extraction. The current application is developed and tested within a Windows environment, while a corresponding web application, which would offer term extraction from texts in various domains to a wider community of expert users, is under development.

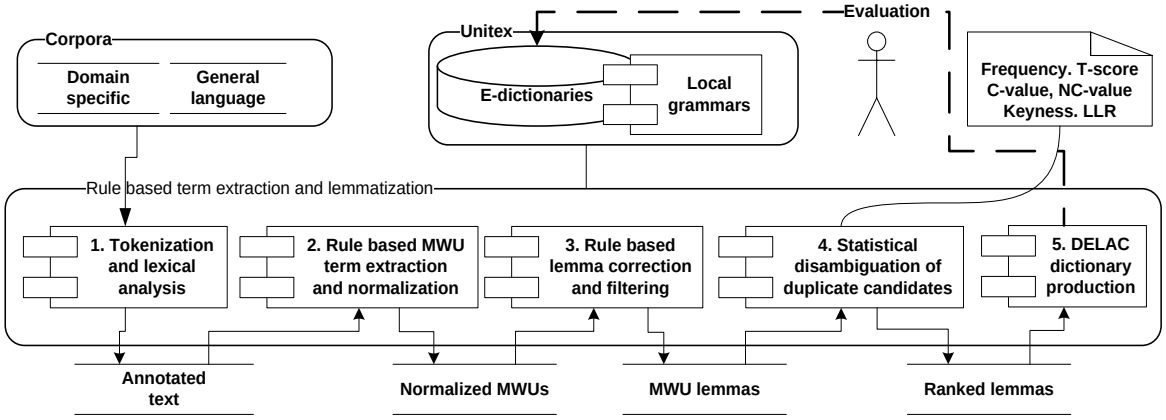


Figure 2. Architecture of the system

4. Evaluation

For evaluation we used a corpus that contains 10textbooks, 2 projects and 51 journal articles from the mining domain. The size of this corpus is 32,633 sentences and 625,105 simple word forms. For calculation of measures that compare results on a domain corpus with general language we used SrpKor.

Our procedure retrieved 85,276 different MWU forms to which lemmas were assigned, resulting in 134,608 candidates where graph, number, lemma and form are taken into account. In the first phase of lemmatization 83,038 different simple word lemmas (*LemmaTemp*) were produced, from which, in the second phase 114,979 MWU lemma candidates were obtained. For evaluation we kept only candidates whose frequency passed the threshold of 7. Distribution of retained MWU forms and lemmas by graphs that retrieved them is given in Table 4.

		sin		plu	
Graph		Lemma	Forms	Lemma	Forms
grf01	AXN	570	11,845	481	8,200
grf02	2XN	6	55	5	53
grf03	N2X	701	11,330	613	6,931
grf04a	N4X	148	2,113	107	1,359
grf04b	N4X	122	2,002	106	1,171
grf05	AXN2X	63	800	33	205
grf06	NXN	299	4,249	134	1,751
grf07	AXAXN	5	37	4	36
grf08a	N6X	6	77	6	46
grf08b	N6X	8	103	8	72
grf08c	N6X	9	103	7	81
grf09a	AXN4X	17	195	10	124
grf09b	AXN4X	9	86	7	47
grf10	2XAXN	8	62	6	100
Total		**1,971**	**33,057**	**1,527**	**20,176**

Total (plu+sin)	**3,498**	**53,233**

Table 4. Number of lemmas that passed the frequency threshold 7

Out of 3,498 retained candidate lemmas, 1,540 were eliminated as false by applying rules described in Section 2, leaving 1,958 lemmas which cover 33,153 forms recognized in the corpus, out of which 4,067 distinct. Distribution of these MWU forms and lemmas by graphs is given in Table 5.

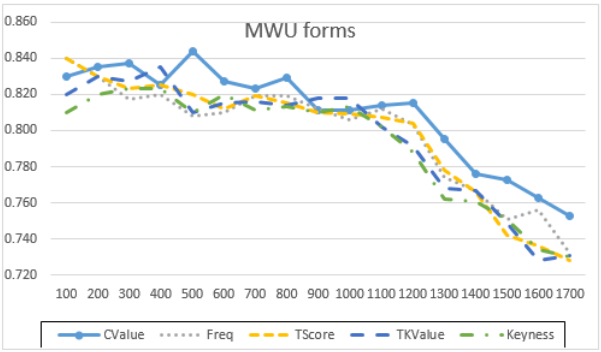
		sin		plu	
Graph		**Lemma**	**Forms**	**Lemma**	**Forms**
grf01	AXN	568	11,828	47	740
grf02	2XN	6	55		
grf03	N2X	668	10,993	8	34
grf04a	N4X	143	2,065	2	22
grf04b	N4X	122	2,002	1	9
grf05	AXN2X	63	800	3	28
grf06	NXN	265	3,903		
grf07	AXAXN	5	37	1	8
grf08a	N6X	6	77		
grf08b	N6X	6	83	1	7
grf08c	N6X	9	103		
grf09a	AXN4X	17	195		
grf09b	AXN4X	8	78		
grf10	2XAXN	8	62	1	24
Total		1,894	32,281	64	872
Total (plu+sin)		**1,958**	**33,153**		

Table 5. Number of lemmas that are passed for manual evaluation

For the forms remaining after automatic filtering measures were calculated, and they were ranked accordingly, as illustrated by Table 3, after which they were passed to the evaluator.

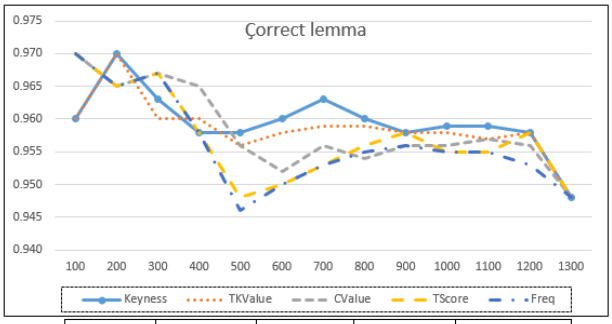
The evaluator performed two tasks: (a) checking for each lemma and all its retrieved forms whether they actually represent a MWU, and (b) verifying for each proposed lemma whether it is a correct lemma. The precision of retrieval was calculated for each of these tasks and the results are presented in Figures 3and 4 for groups of hundreds ranked by basic measures: Frequency, C-Value, T-Score, and Keyness and one combined measure TKValue = T-Score * Keyness. Mean average precision given at the bottom of Figure 2 shows that all measures gave comparable results.

The evaluation also showed that our extraction graphs missed some term structures as well as some terms having more than four components. For instance, the graphs recognized as terms *motor mehanizma za potiskivanje* (structure N6X) 'pusher mechanism engine' and *mehanizam za potiskivanje kašike* (structure N6X) 'bucket pusher mechanism' while the complete correct term has five components *motor mehanizma za potiskivanje kašike* (structure N8X) 'bucket pusher mechanism engine'.



0.804	0.794	0.793	0.792	0.789

Figure 3. Precision of retrieval by MWU forms



0.960	0.959	0.958	0.957	0.956

Figure 4. Precision of lemmaproduction

Out of 4067 distinct forms, 3836 (94%) were evaluated as proper MWUs and 231 (6%) were removed as not being proper MWUs. Among proper MWUs there were 3715 (97%) with a correct lemma and 121 (3%) with an incorrect lemma (Figure 5).

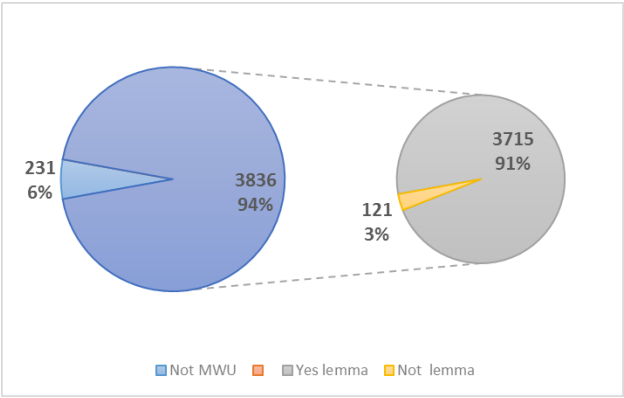


Figure 5. Evaluation of MWU forms and correct lemma

			MWU OK		Lemma OK	
		Graph	No	%	No	%
plu	grf01	AXN	42	89	8	19
	grf03	N2X	7	88	3	43
	grf04a	N4X	1	50		0
	grf04b	N4X	1	100		0
	grf05	AXN2X	3	100		0
	grf07	AXAXN	0	0		0
	grf08b	N6X	1	100		0
	grf10	2XAXN	1	100		0
sin	grf01	AXN	553	97	552	100
	grf02	2XN	6	100	5	83
	grf03	N2X	608	91	599	99
	grf04a	N4X	102	71	97	95
	grf04b	N4X	111	91	106	95
	grf05	AXN2X	58	92	58	100
	grf06	NXN	29	11	13	45
	grf07	AXAXN	4	80	4	100
	grf08a	N6X	5	83	5	100
	grf08b	N6X	6	100	5	83
	grf08c	N6X	7	78	7	100
	grf09a	AXN4X	10	59	10	100
	grf09b	AXN4X	7	88	7	100
	grf10	2XAXN	8	100	8	100
	Total		1570	83	1487	95

Table 6 Number of lemmas after manual evaluation

5. Concluding remarks and Future Work

The paper presents an approach to terminology extraction for Serbian based on e-dictionaries and local grammars. For extraction purposes 14 graphs were developed, which extract the most frequent syntactic structures identified by an analysis of several Serbian terminological dictionaries and Serbian e-dictionary of MWUs. The approach was evaluated on the example of terminology extraction from a mining corpus, with results for extraction, normalization and lemmatization presented successively. Automatic generation of a complete lemma in the form required by the electronic dictionary of Serbian language is a challenging task, which has previously not been tackled, and thus presents the most important contribution of this paper.

When determining the lemma, in the first phase a number of possible candidates are generated, from which, automatically, using a system of rules and the frequency of singular and plural forms, one of the possible lemmas is selected. Finally, a detailed evaluation of the results was performed manually, as presented by tables and graphs in Section 4.

The solution to terminology extraction outlined in this paper will by all means speed up the development of e-dictionaries, as in addition to the terminology extraction, the approach can be applied to the extraction of MWUs belonging to general lexica. Expanding the e-dictionaries will further improve systems for information retrieval, information extraction, query expansion and the like. One useful application can also be the creation of bilingual and multilingual terminological dictionaries, which would provide coverage of terms from a specific domain.

In our future work we will concentrate on:

- Finalization of the web application;
- Improvement of the precision of correct lemma production (development of additional strategies to avoid offering of incorrect lemmas). For instance, our results showed that at the end we obtained just a few (11) lemmas in plural (see Table 6), all of them recognized by just two graphs. This suggests that lemmas in the plural form that prevail among those that were offered and then rejected should be offered for a limited number of structures, e.g. if recognized only by one of these two graphs.
- Development of new extraction FSTs for additional syntactic structures of MWTs, especially for terms with more than four components;
- Application to various different domains (information and library sciences, electro-energetics, etc.);
- Experiments with different strategies and measures for distinguishing general-language MWUs from domain-specific MWTs.

6. Acknowledgements

This research was supported by the Serbian Ministry of Education and Science under the grant #47003 and #178003. The authors would also like to thank the anonymous reviewers for their helpful and constructive comments.

7. Bibliographical References

Ammar, C., Haddar, K., and Romary, L. (2015). Automatic Construction of a TMF Terminological Database Using a Transducer Cascade. In *Proc. of RANLP*, pp. 17--23.

Bounhas, I., and Slimani Y. (2009). A hybrid approach for Arabic multi-word term extraction. *Proc. Of the Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009*, pp. 1--8.

Broda, B., Derwojedowa, M., and Piasecki, M. (2008). Recognition of structured collocations in an inflective language. *Systems Science*, *34*, pp. 27--36.

Chen, J., Yeh, C. H., and Chau, R. (2006). A multi-word term extraction system. *PRICAI 2006: Trends in Artificial Intelligence*, Springer Berlin Heidelberg, pp. 1160--1165.

Church, K. W., Hanks, P., (1990). Word association norms, mutual information, and lexicography, *Computational Linguistics*, 16, pp. 22--29.

Church, K. W. Gale, W., Hanks, P., Hindle, D. (1991). *Using statistics in lexical analysis*, In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 115--164.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), pp.7--36.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), pp.61--74.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:.the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), pp. 115-130.

Gelbukh, A., Sidorov, G., Lavin-Villa, E., and Chanona-Hernandez, L. (2010). Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference. In: Hopfe, C. J., Rezgui, Y., Métais, E., Preece, A., Li, H. (Eds.), *Natural Language Processing and Information Systems.* Berlin: Springer, pp. 248--255.

Koeva, S. (2007). Multi-word term extraction for Bulgarian. In *Proc. of the Workshop on BSNLP: Information Extraction and Enabling Technologies*, pp. 59--66.

Krstev, C., Obradović, I., Stanković, R., and Vitas, D. (2013). An Approach to Efficient Processing of Multi-Word Units. In: Przepiórkowski, A., Piasecki, M., Jassem, K., Fuglewicz, P. (Eds.) *Computational Linguistics*. Berlin: Springer, pp. 109--129.

Krstev, C., Stanković R., Obradović I., and Lazić B. (2015). Terminology acquisition and description using lexical resources and local grammars. In *Proc. of the Conf.Terminology and Artificial Intelligence 2015,* Granada: University of Granada, pp. 81--89.

Malyszko, J., Abramowicz, W., Filipowska, A., & Wagner, T. (2015). Lemmatization of Multi-Word Entity Named for Polish Language Using Rules Automatically Generated Based on the Corpus Analysis. In *Proc. of 7th Language & Technology Conference* 2015, Poznań: Fundacja Uniwersytetuim. A. Mickiewicza, pp. 540--544.

Pantel, P., Dekang L. (2001). A statistical corpus-based term extractor. In Stroulia, E., Matwin, S. (Eds.) *Advances in Artificial Intelligence*, Berlin: Springer Berlin Heidelberg, pp. 36--46.

Piskorski, J., Sydow, M., and Kupść, A. (2007). Lemmatization of Polish person names. In *Proc. of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, Stroudsburg: Association for Computational Linguistics, pp. 27--34.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek A., and Makowiecki F. (2012). SEJFEK — a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proc. of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012*, Mumbai*: COLING*, pp. 195--214.

Schone, P., Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proc. of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning -Volume 7*, Stroudsburg: Association for Computational Linguistics, pp. 67--72.

Smadja, F.(1993). Retrieving Collocations from Text: Xtract, *Computational Linguistics*, 19(1), pp. 143-177.

Stanković, R., Obradović, I., Krstev, C., and Vitas, D. (2011). Production of morphological dictionaries of multi-word units using a multipurpose tool. In *Proc. of the Computational Linguistics-Applications Conference, October 17-19, 2011*, Jachranka: Polskie Towarzystwo Informatyczne, pp. 77--84.

Tadić, M., Šojat, K. (2003). Finding multiword term candidates in Croatian. In *Proc. of IESL2003 Workshop, Borovets: Context*, pp. 102-107.

Vintar, Š. (2010). Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology, 16*(2), pp.141--158.

Vitas, D., Popović, Lj., Krstev, C., Obradović, I., Pavlović-Lažetić, G. and Stanojević, M. (2012). The Serbian Language in the Digital Age. Berlin; Springer-Verlag.

8. Language Resource References

Vitas D., Utvić M. (2015). SrpKor22M, Serbian automatically lemmatized, PoS and morphosyntactically annotated corpus 22M words, not disambiguated, distributed via owner.

Stanković R. (2015). LeXimir v2, tool for creation, management and exploitation of lexical resources, distributed via owner.

Krstev C., Vitas D. (2015). SrpRec, Serbian morphological electronic dictionary, http://www-igm.univ-mlv.fr/~unitex/index.php?page=5, CC BY-NC-ND.

Lazić B., Stanković R. (2015), MineCorp, Serbian corpus from mining domain.