Semi-Automatic Extraction of Multiword Terms from Domain-Specific Corpora

Vesna Pajić, Staša Vujičić Stanković, Ranka Stanković, Miloš Pajić

Semi-Automatic Extraction of Multiword Terms from Domain-Specific Corpora | Vesna Pajić, Staša Vujičić Stanković, Ranka Stanković, Miloš Pajić | The Electronic Library | 2018 | 36 | 3

10.1108/EL-06-2017-0128

emerald PUBLISHING

The Electronic Library

Semi-automatic extraction of multiword terms from domain-specific corpora

Journal:	*The Electronic Library*
Manuscript ID	EL-06-2017-0128.R2
Manuscript Type:	Article
Keywords:	Data analysis, Data processing, Data retrieval, Digital documents, Document handling, Evaluation, Foreign languages, Information retrieval

Semi-Automatic Extraction of Multiword Terms from Domain-Specific Corpora

1 Introduction

Multiword expressions (MWEs) are lexical units composed of more than one word, which are syntactically, semantically, pragmatically and/or statistically idiosyncratic (Baldwin and Kim (2010)). Domain-specific MWEs are usually referred to as "multiword terms" (MWTs). It is estimated that they constitute a significant portion of terminology; over 70% of the terms are complex lexical units (da Graça Krieger and Finatto (2004)). It is difficult to identify them automatically using existing methods, because there are relatively few MWT instances in big corpora which cannot be spotted by exploiting their statistical properties only. The extraction is even more complex for languages with rich morphological system, such as Serbian (Mariani (2005), Vitas *et al*. (2005)).

Collection and extraction of MWTs are two of the most important steps in the process of creating a terminological lexicon and they are also the most time-consuming. Human expert engagement cannot and should not be avoided, but such work could be significantly facilitated by well-designed automatic or semi-automatic extraction procedures. Thus we focus on developing a method for identifying and extracting MWTs directly from domain-specific corpora, which is suitable for processing morphologically rich languages.

In this paper, we present a hybrid approach that combines linguistic and statistical information to extract term candidates from texts in languages with rich morphological system. The method is designed to be domain and language independent, although we focus on identification and extraction of MWTs from texts in Serbian belonging to the domain of agricultural engineering, as a use case.

2 Related work

In the last two decades, there has been considerable NLP research into MWEs (Liang et al. 2017b; Nakov and Hearst, 2013; Ramisch, 2015; Tsvetkov and Wintner, 2014). The work on MWEs in English still dominates, although there has been some research in languages other than English: Czerepowicka and Savary (2015) for Polish; Liang et al. (2017a) for Chinese; Macken and Tezcan (2016) for Dutch; Mandravickaite and Krilavičius (2017) for Latvian and Lithuanian; Zaninello and Nissim (2010) for Italian and more.

The problem of MWE extraction from literary texts in Serbian was described in detail in (Krstev et al. 2014). The authors presented finite state automata (FSA) for describing MWEs that have a predictable structure and potentially infinite number of instances (e.g. date and time expressions). They also identified the most frequent structures of Serbian MWEs. These structures will be further explained in Section 3, since we used them for extracting MWTs in our experiment.

Automatic term extraction is an important part of NLP systems. It is used for lexicon creation, acquisition of novel terms, text classification, text indexing, machine-assisted translation and other NLP tasks. Different approaches to MWT extraction, linguistics-based or statistics-based (or both), have already been published recently (Cram and Daille (2016), Sclano and Velardi (2007), Verberne et al. (2016), Vivaldi and Rodríguez (2007), Yin et al. (2016), Zhang and Wu (2012)). Most of the methods used for MWT extraction today are hybrid, i.e. they usually integrate statistical information, such as frequencies of n-grams and collocations, with linguistic information, such as syntactic patterns of expressions. There is no consensus on the best method, or even if there is any. It depends on what expressions are considered to be MWTs, and also on the level of their compositionality, the text domain, language specifics and application needs. As statistical information, different frequency and association measures are being used in the MWT extraction process, such as T-score (Church

et al. (1991)), the log-likelihood ratio (LLR) (Dunning (1993)), C/NC value (Frantzi *et al.* (1998)), Keyness (Scott and Tribble (2006)), and others.

The multifaceted problem of terminology also attracted significant attention from researchers in Slavic languages who approached it from various perspectives: extraction, description, multilinguality. For terminology extraction, the variety of mentioned approaches were used (Tadić and Šojat (2003), Vintar (2004), Koeva (2007), Savary and Zaborowski (2012), Przepiórkowski *et al.* (2007)).

3 Acquisition of multiword terms from the agricultural domain in Serbian

3.1 The main objectives and context of the work

The main objectives of the research include: (i) creation of an up-to-date terminological lexicon for agricultural engineering in Serbian; (ii) expansion of the existing Serbian morphological dictionary of compounds; (iii) MWT structure analysis in Serbian, in order to improve the methods of their automatic acquisition in the future, and (iv) automation of the MWT acquisition process to some extent. Here, we present work aimed at contributing to the achievement of objective (i), i.e. achieving objectives (ii) to (iv), with experiments done on the texts from the agricultural engineering domain in Serbian. The same approach can be applied to other domains and languages.

3.2 Some specifics of the Serbian language

Serbian belongs to the group of Slavic languages and it has a rich morphological system. For instance, nouns inflect for number and case, while adjectives inflect for number, gender, case, and degree. This abundance of forms requires specific treatment.

The problem can be illustrated by the example of the noun phrase *jeftina radna snaga* – the Serbian expression for ‘cheap labor’ (literally ‘cheap labor force’). It is a multiword noun that inherits its gender from the constituent noun *snaga* ‘force’ (which is feminine), and

it inflects for case, but it is not inflected for number (although the simple word *snaga* is). The adjectives *jeftina* 'cheap' and *radna* 'labor' agree with the noun *snaga* in number, case, gender and animacy, but the comparative and positive forms of these adjectives are not used in this noun phrase. When inflected separately, the adjective *jeftina* has 204 different grammatical forms, *radna* has 76, and *snaga* has 7, which creates a total of 108,528 different combinations of forms for the whole text sequence[1]. However, only 7 of them are grammatically correct in Serbian.

As a consequence, NLP methods developed for the English language cannot be applied to Serbian texts with the same precision and efficiency. Use of electronic resources, such as lexicons, grammars and dictionaries is indispensable in order to process texts in the Serbian language.

3.3 Syntactic patterns

Prior to this research, agricultural terminology in Serbian had not been studied from the perspective of computational linguistics. We had no *a priori* knowledge of the syntactic structure of MWTs belonging to this domain. Instead, we used syntactic patterns of MWEs in Serbian, based on previous analyses of their structures (Krstev *et al.* (2013), Stanković *et al.* (2011), Utvić, M. (2011)). From the terminology perspective, we were interested in nouns (used to name concepts).

The complete list of syntactic patterns used in our research is given in Table I. The choice of syntactic patterns was based on previous analyses of the structures of the MWEs (Krstev et al. 2013), already present in Serbian e-dictionaries (that contained some

1 In many cases a word form has several grammatical forms. For instance, the word form *radnu* has 5 grammatical forms: a definitive masculine gender form in the dative case singular, a feminine gender form in the accusativ singular, etc.

terminology mostly from library and information sciences) (see Section 3.6). These analyses revealed that the most frequent nominal MWEs in Serbian have two components (82.9%), followed by MWEs with three components (13.7%), while longer MWEs are much less frequent (3.4%).

Table I Syntactic patterns used for MWT extraction.

ID	Syntactic pattern*	Description	Example
1	AN	Both components inflect and agree in gender, number and case	*žitni kombajn* 'wheat harvester'
2	X-N	The first component does not inflect (and is never used as an independent word in Serbian); the second component inflects	*eko-sistem* 'ecosystem'
3	NNgi	Only the first component inflects; the second is always in genitive	*obrada zemljišta* 'soil tillage', lit. 'tillage of soil'
4a	NPrepNp	Only the first component inflects, followed by prepositional phrase	*sistem za navodnjavanje* 'irrigation system', lit. 'system for irrigation'
4b	NNgiNgi	The first component inflects; the second and third are always in genitive	*tok vegetacionog perioda* 'course of vegetation period'
5	ANNgi	The first and the second components inflect, the third is always in genitive	*unutrašnji deo parcele* 'inner part of a parcel'
6	N(-)N	Both components inflect and agree in case and number; can be optionally separated by a hyphen	*traktor guseničar* 'tractor caterpillar'
7	AAN	All components inflect and agree in gender, number and case	*elektronska komandna jedinica* 'electronic command unit'
8a	NNgiPrepNp	Only the first components inflects; the second is always in genitive, followed by prepositional phrase	*sistem mašina za doradu* 'machine system for final treatment'
8b	NNgiNgiNgi	Only the first components inflects; the rest are always in genitive	*redukcija stepena zagađenja vazduha* lit. 'reduction of the level of the pollution of air'
8c	NPrepNpNgi	The first component inflects;it is followed by prepositional phrase, the fourth is in genitive	*mašina za obradu zemljišta* lit. 'machine for tillage of soil'
9a	ANPrepNp	The first two components inflects and agree in case, gender and number; the last component is prepositional phrase	*tehnički sistemi za redukciju* 'technical systems for reduction'

9b	ANNgiNgi	The first two components inflects and agree in case, gender and number; the last two components are always in genitive	*energetska upotreba drvne biomase* lit. 'energetic utilization of tree biomass'
10	X-AN	The first component does not inflect (and is never used as an independent word in Serbian); the second and third component inflect and agree in case, number and gender	*tehničko-tehnološko rešenje* lit. 'technical-technological solution'

* A – adjective, N – noun, Ngi – noun or adjective in the genitive or instrumental case, "-" – the separating hyphen, (-) – optional occurrence of the separating hyphen, PrepNp – prepositional phrase, X – invariable morpheme.

3.4 Measures used for term candidates

Usage of common association measures of unithood and termhood, such as T-Score, LLR, C-Value and the like, requires complex text preprocessing (lemmatization or normalization) on the level of a corpus. Such preprocessing of morphologically rich languages is error-prone, which makes calculating those measures (or some of their parts) difficult, sometimes even impossible.

Instead of modifying the existing association measures or doing time-consuming and error-prone lemmatization of the whole corpus, we chose to use frequencies of occurrence of a text sequence in the corpus, combined with normalization described later in Section 3.7. This approach was justified by the obtained results, since it remained relatively simple, but achieved high precision, compared to other methods (see Section 4.1).

3.5 Corpus

The collection of texts used for our research consists of scientific papers from the domain of agricultural engineering, written in Serbian. The corpus has over 621,874 simple word forms (out of which 42,262 are unique word types), recognized by Serbian electronic dictionary (see 3.6), and 8,457 unknown simple word forms (word forms not listed in dictionaries). Additionally, 28,698 MWEs were identified (as already included in Serbian

morphological dictionaries). Among compound entries, only 470 were adjective or noun phrases forms, while the rest were multiword numerals.

The number of unrecognized simple words and multiword numerals was higher than in some other Serbian corpora, since this corpus was created automatically, from journal files in PDF format. During the conversion from PDF to TXT files, the tables from the original text were converted to numerical text sequences. In addition, there were a lot of English words (obtained from some references, for example), errata, and misspelled words.

3.6 Language resources and software tools

For linguistic corpus analysis and pattern search, we used the Unitex software system[2]. It is a corpus processing system, based on an automata-oriented technology. Unitex enables application of morphological electronic dictionaries and grammars to texts in a number of different languages for different kinds of natural language morphological, syntactic, and semantic processing (Courtois *et al*. (1990)). These dictionaries are created in plain text format, where each line contains a word entry, i.e. its inflected form, the lemma of the word and various grammatical morphosyntactic, semantic and other information (DELAS, DELAF, DELAC, and DELACF formats (Savary (2008))).

The morphological dictionary of MWEs for Serbian is in DELAC format and has 13,676 entries (Krstev et al. 2013). It contains both general lexica and proper names. For example, the entry in the DELAC dictionary of Serbian for the MWE *jeftina radna snaga (eng. cheap labor force)* is:

```
jeftina(jeftin.A17:aefs1g)
radna(radni.A2:aefs1g)snaga(snaga.N610:fs1q),NC_AXAXN1+HumColl
```

2 The Unitex software system: http://www-igm.univ-mlv.fr/~unitex/.

The information given in this entry, together with the inflectional transducer NC_AXAXN1, allows automatic production of all 7 correct inflected forms. Every generated form has assigned codes of the values of grammatical categories, as well as markers (inherited from the corresponding lemma) that describe its semantic, dialectical, domain or other features (e.g. in the given example the marker +HumColl stands for 'human collective'). For example, one generated form (the dative case), from the DELACF dictionary, of the MWE *jeftina radna snaga* is:

```
jeftinoj radnoj snazi,jeftina radna snaga.N+HumColl:fs3q
```

In Unitex, dictionaries and grammars are applied to text through graphs corresponding to finite state automata (FSA) and finite state transducers (FST). Figure 2 represent one such graph.

3.7 Steps in term extraction and dictionary production

One of the main objectives of our research was to expand the Serbian e-dictionary with new entries automatically. We concentrated on texts from the agricultural engineering domain and on extracting MWTs from them. Krstev et al. (2013) already automated the production of complex dictionary entries in DELAC format for a given list of MWEs, using e-dictionaries of Serbian simple words, inflectional FSTs and a set of grammatical rules. Therefore, we focused on identifying MWTs in a text, in a manner similar to that proposed by (Krstev et al. 2014), but modified to suit the patterns explained in Section 3.3.

The extraction of MWTs from the text was performed in several steps (Figure 1).

STEP 1: CORPUS PREPARATION. The text was converted automatically from other document formats (.PDF, .DOC etc.) to plain text format using a specific software tool developed by the authors (Pajić *et al.* (2012)).

STEP 2: LINGUISTIC PRE-PROCESSING. In this step we tokenize the corpus and split it into sentences before applying dictionaries to it, building the subset of dictionaries consisting only of the forms that were present in the corpus (e.g. `agrarna,agrarni.A+PosQ:akms2g`). This subset is called *the dictionary of the text*.

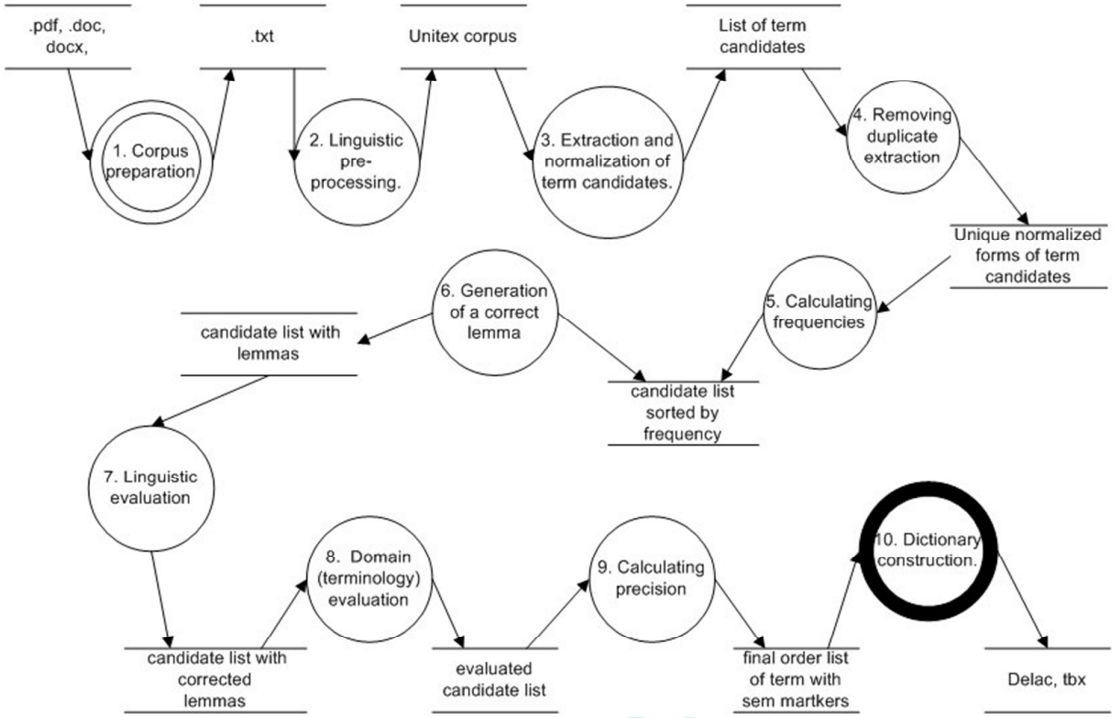

Figure 1. *Steps in term extraction and dictionary production.*

STEP 3: EXTRACTION AND NORMALIZATION OF TERM CANDIDATES. In order to avoid recognition of different inflectional forms of the same lexical units as different MWTs, we need to somehow normalize or lemmatize the text. As stated in Section 3.2, the lemmatization of MWEs in Serbian is a difficult task that cannot be done correctly without linguistic information about an MWE, as given in DELAC entries (Section 3.7). Since we did not have this type of information for novel agricultural domain-specific terms found in the corpus, we chose to perform normalization of term candidates by lemmatizing word by word (the first round of normalization). This kind of normalization is done together with term candidate extraction from the text. For this task, we designed several FSTs that (i)

recognize syntactic structures from Table I in the corpus, and (ii) extract term candidate lemmatized word-by-word. Recognition, extraction, and lemmatization are done with a single graph (transducer) or with one network of graphs (transducers) per syntactic pattern. Figure 2 shows FSTs for detecting NNgi pattern.

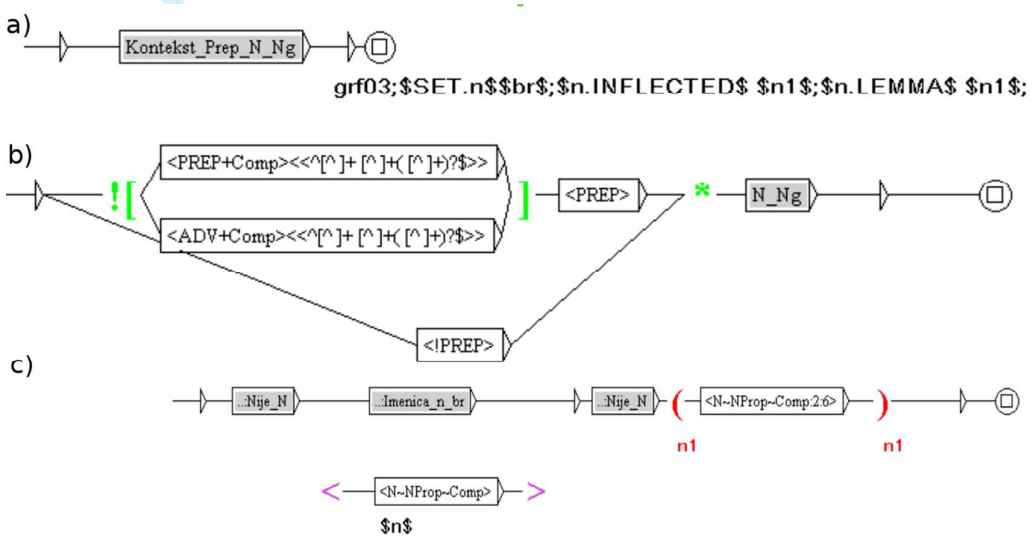

Figure 2. *Graph network for retrieving text sequences conforming to the syntactic pattern NNgi. a) Top level graph produces the final output based on subgraph Kontekst_Prep_N_Ng, b) Graph Kontekst_Prep_N_Ng which analyzes the context of NNgi sequence, c) Graph N_Ng for detecting particular MWT – a noun followed by a noun that does not inflect in the MWT (usually in the genitive or instrumental case – "2:**6**").*

The graphs were applied one by one and their results (the output of transducers) were merged into dataset and stored in a database for further processing. In this phase, for each term candidate we tracked its form found in the text, its suggested normalized form (consisting of simple word lemmas) and the label of the graph(s) that extracted it. An excerpt of the resulting table from this phase is given in Table II (the column with the English translation is added here for the sake of clarity).

Table II *Extracted term candidates, with their normalized forms and graph labels.*

Term candidate occurrence	Normalized form (lemmatized word by word)	Graph pattern	English translation
agrarne politike agrarnom politikom agrarnu politiku	agrarni politika	AN	agricultural policy
merač pritiska	merač pritiska	NNgi	pressure gauge
merača pritiska	merač pritiska	NNgi	pressure gauge

STEP 4: REMOVING DUPLICATE EXTRACTION. Some forms were recognized by two or more graphs. For example, *rezultati istraživanja* 'results of research' were correctly recognized as NNgi, and falsely as N-N, due to homography of the form *istraživanja* (Table III). In this phase, we automatically removed those duplicate values, leaving only the candidates that were extracted by using the graph for more frequently used structures. In this case, the structure NNgi is much more frequent than N-N.

Table III *An excerpt from the table with numbers of occurrence of term candidates, grouped by normalized form and graph. It can be seen that the two occurrences were recognized by using two different graphs, producing different normalized forms.*

Graph	Number	Word form	Temporary lemma	Lemma	Frequency
NNgi	plu	rezultata istraživanja	rezultat istraživanja	rezultati istraživanja	31
		rezultate istraživanja			2
		rezultati istraživanja			53
		rezultatima istraživanja			30
	sin	rezultat istraživanja		rezultat istraživanja	6
		rezultata istraživanja			31
		rezultate istraživanja			2
		rezultatu istraživanja			1
N-N	sin	rezultata istraživanja	rezultat istraživanje	rezultat istraživanje	31

STEP 5: CALCULATING FREQUENCIES. Unique normalized forms of term candidates were counted and their frequencies saved and analyzed further. Since we assumed that the most frequent candidates would be evaluated as terms, the list of candidates was sorted in descending order, with high frequency terms at the top.

STEP 6: GENERATION OF A CORRECT LEMMA. The term candidates that we obtained after the first round of normalization were not always the correct MWE lemmas which depended on their syntactic structure, and consequently, the FST that was used for their extraction (the problem was with graph AN, because adjective did not always agree in gender with a corresponding noun, as it should). The second round of normalization consisted of correcting these lemmas. For that purpose we used the same tools and resources: Unitex, e-

dictionaries and FSTs, only this time we used "inverted dictionaries" along with regular ones. In these dictionaries the form of entries was not as usual `form,lemma.POS+Markers:gram_cat` but `lemma,form.POS+Markers:gram_cat`. The correcting FSTs read the list of term candidates prepared in previous steps and produced the form of an adjective that agreed in gender with a corresponding noun.

STEP 7: LINGUISTIC EVALUATION. Extracted term candidates and their lemmas produced as described in this section were evaluated by a linguist first and corrected manually, if needed.

STEP 8: DOMAIN (TERMINOLOGY) EVALUATION. The extracted term candidates were further estimated by two human experts in the field of agricultural engineering, whether they are MWT from the domain or not. Agricultural engineering is part of a broader agricultural domain and has a lot of terms in common with other technical domains (such as civil engineering). That is why we used two domain categories: *agricultural term* and *technical term*. Each extracted term candidate was labelled with *yes* or *no* depending on whether it belonged to those categories or not. In that way we could subsequently select the sequences with both categories set to *yes* as MWTs from the agricultural engineering domain. By contrast, sequences with both categories set to *no* were of no interest to the agricultural engineering domain. They were either MWTs from other domains, such as *najveći broj* 'maximum number' or some errata and wrongly extracted words. The evaluation process and results are described in more detail in Section 4.

STEP 9: CALCULATING PRECISION. Since we were interested in the estimation of the frequency value that can be used as a cut-off value for automatic extraction, we used precision at rank n ($P@n$) measure on a sorted list of candidates. Let $e[i]$, $i=0 \ldots n$ be the sorted array of top n extracted terms, where $e[0]$ is the most frequent sequence. We define the corresponding frequencies as $freq(i)$. Let $term(n)$ be the number of sequences

categorized as terms in the set of the first *n* top frequent sequences. For every *n* we calculate precision at rank *n* as $P@n = Prec(n) = term(n)/n$.

STEP 10: DELAC CONSTRUCTION. As a final step, term lemmas were prepared for insertion into the DELAC dictionary of Serbian (Krstev *et al*. (2013)). For each term evaluated as belonging to one of the categories, the appropriate semantic markers were added (+DOM=Agr for agricultural, +DOM=Tech for technical). Some examples of entries from the final term list are presented in Table IV.

Table IV *An excerpt from the list of terms with semantic markers.*

MWT	Semantic markers	English translation
brzina kretanja	+DOM=*Tech*	moving speed
izvor energije	+DOM=*Agr*+DOM=*Tech*	energy source
poljoprivredna proizvodnja	+DOM=*Agr*	agricultural production

The corrected lemmas were further processed by using LeXimir (Obradović and Stanković (2008)), which automatically produced the following illustrative DELAC entries (but also can produce a TermBase eXchange (TBX) format (Romary (2014))):

```
brzina kretanja(kretanje.N300:np1q),NC_2XN3+DOM=Tech
izvor(izvor.N1:ms1q) energije,NC_N2X+DOM=Agr+DOM=Tech
poljoprivredna(poljoprivredni.A2:aefs1g)
proizvodnja(proizvodnja.N660:fs1q),NC_AXN+DOM=Agr
```

4 Results and discussion

During the research, more than 50,000 occurrences of different forms of possible MWTs were extracted. After normalization step, because of large number of candidates for manual evaluation, we chose 1,523 most frequent term candidates (with frequency higher than 8), which covered 22,579 occurrences in text. The maximum frequency a term candidate had was 346. Among them, only 58 were already included in the DELAC dictionary of Serbian.

When grouped by syntactic structures, the results were as shown in Table V.

Table V *The results of extraction of term candidates after the STEPS 3 - 7.*

Pattern	Term candidates	Evaluated as MWT	Correct lemma
AAN	1	1	1
AN	417	406	386
ANNgi	74	59	67
ANNgiNgi	13	8	12
ANPrepNp	20	12	14
N-N	158	66	10
NNgi	526	459	476
NNgiNgi	130	106	114
NNgiNgiNgi	9	7	9
NNgiPrepNp	11	9	11
NprepNp	137	87	109
NPrepNpNgi	20	17	17
X-AN	6	4	3
X-N	1	1	1
Total	**1523**	**1242**	**1230**

After the evaluation, if any of the evaluators assessed that a term candidate was a term from any of the two categories, we marked this candidate as a term (column IT ('Is Term') in Table VI). The precision at rank n (the last column of Table VI) was calculated based on the values from the column IT ('Is Term').

Table VI *An excerpt from the resulting table with the most frequent terms and their evaluation.*

Term candidate	Evaluator 1		Evaluator 2		Frequency	IT	P@n
	Agr	Tech	Agr	Tech			
električna energija 'electricity'	no	yes	no	yes	346	yes	1.00
brzina kretanja 'moving speed'	no	yes	no	yes	320	yes	1.00
poljoprivredna proizvodnja 'agricultural production'	yes	no	yes	no	211	yes	1.00
obrada zemljišta 'soil tillage'	yes	yes	yes	yes	200	yes	1.00
radna brzina 'working speed'	no	yes	no	yes	196	yes	1.00
potrošnja energije 'energy consumption'	no	yes	no	yes	195	yes	1.00
energetska efikasnost 'energetical efficiency'	no	yes	no	yes	180	yes	1.00
poljoprivredna mehanizacija 'agricultural mechanization'	yes	yes	yes	yes	146	yes	1.00
snaga motora 'power of engine'	no	yes	no	yes	143	yes	1.00

sistem mašina 'machine system'	no	yes	no	yes	133	yes	1.00

Evaluators' agreement was measured with the Cohen's kappa coefficient (Viera *et al.* (2005)) and it showed a substantial agreement between our two evaluators ($0.61 < \kappa < 0.80$) (Table VII).

Table VII *Evaluator's agreement.*

	Agr			Tech		
	Eval(1)	**Eval(2)**	**Agree**	**Eval(1)**	**Eval(2)**	**Agree**
ny*	1053	1076	486	1103	1094	461
nn	587	564	975	537	546	1018
$Total$	1640	1640	1461	1640	1640	1479
$p(y)$	0.642	0.656		0.673	0.667	
$p(n)$	0.358	0.344		0.327	0.333	
$Pr(a)$			0.891			0.902
$Pr(e)$			0.544			0.558
κ			0.760			0.778

* ***ny*** – number of 'yes' answers, ***nn*** – number of 'no' answers.
Total – total number of answers (total number of answers where both evaluators agree).
p(y) – probability of a 'yes' answer, $p(n)$ – probability of a 'no' answer.
Pr(a) – probability of an answer where both evaluators agree,
Pr(e) $= p_1(y) * p_2(y) + p_1(n) * p_2(n)$, $\boldsymbol{\kappa} = (Pr(a) - Pr(e))/1 - Pr(e)$.

In this way, among the first 1,523 term candidates we had 928 evaluated as terms, 870 of which were new, not already included in the DELAC dictionary of Serbian, so we expanded it with these new terms. The entries of the 58 terms already contained in the dictionary were extended with labels +DOM=Agr and/or +DOM=Tech, depending on the evaluators' opinion.

Precision at rank n gave us information on how many of the first n entries were qualified as terms. One of the objectives of this research was to determine a cut-off value for frequency that can be used for automatic extraction of terms in some future applications, and that is why it was important to analyze the relation between the rank of term candidate and the precision achieved in the set of candidates. The precision at rank n for all candidates is given in Figure 3.

In Figure 4 we showed *P@n* for the top 100 term candidates only. Information about how precision changed with different frequencies (Table VIII) can be used for estimating a cut-off frequency value for achieving a particular precision.

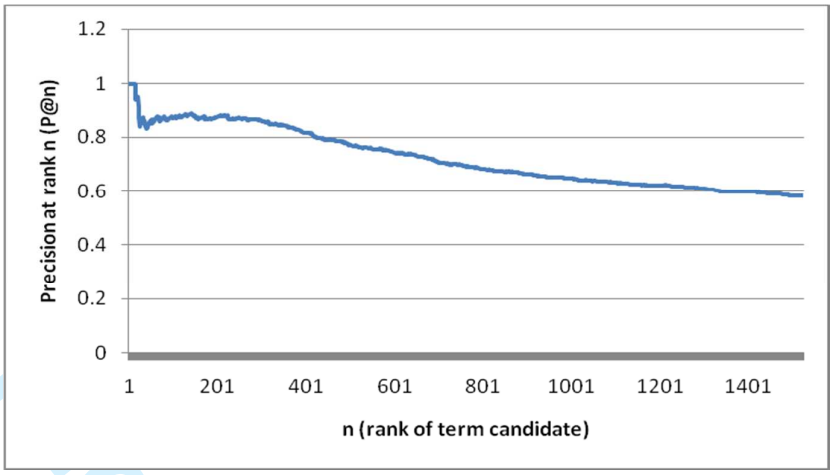

Figure 3. *Precision at rank n for all evaluated term candidates.*

The structures AN and NNgi extracted many more terms than others (Table IX). The graph for extracting terms having the AN structure had a higher precision (81%) than the NNgi graph (61%).

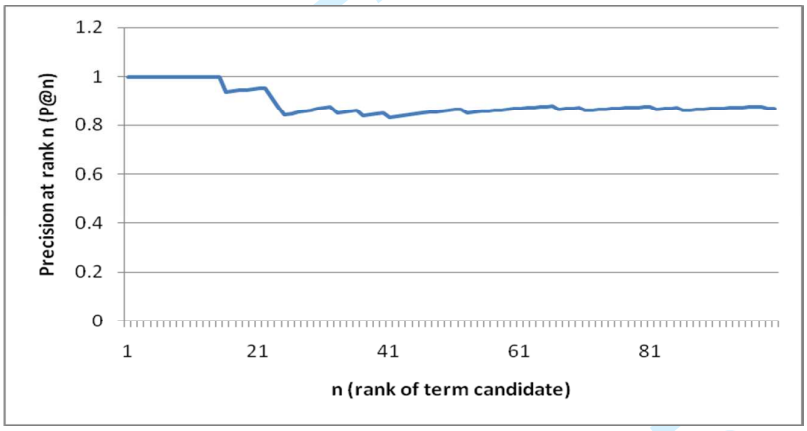

Figure 4. *Precision at rank n for the top 100 evaluated term candidates.*

Table VIII *The frequencies and precision for first n term candidates.*

n	*freq(n)*	*prec(n) = P@n*	*n*	*freq(n)*	*prec(n) = P@n*
1	346	1	800	14	0.68
100	48	0.87	900	12	0.66
200	33	0.87	1,000	12	0.65
300	26	0.86	1,100	11	0.63
400	21	0.81	1,200	10	0.62
500	18	0.77	1,300	10	0.61
600	16	0.74	1,400	10	0.60
700	15	0.71	1,500	9	0.58

Table IX *The number of terms extracted by particular graphs, i.e. having particular structures (out of 1,744 evaluated terms).*

Syntactic structure	Total number of term candidates (n_{tc})	Total number of terms (n_t)	n_t/n_{tc} ratio
AAN	1	1	1.00
AN	469	380	0.81
ANNgi	72	39	0.54
ANNgiNgi	9	1	0.11
ANPrepNp	22	5	0.23
N-N	181	70	0.39
NNgi	493	300	0.61
NNgiNgi	115	51	0.44
NNgiNgiNgi	7	1	0.14
NNgiPrepNp	10	2	0.20
NprepNp	120	29	0.24
NPrepNpNgi	16	3	0.19
X-AN	7	5	0.71
X-N	1	0	0.00

The recall was estimated based on manual extraction of terms from the randomly selected paragraphs containing 2,500 words. The evaluators extracted 115 multi-word terms from the text, with 67 of them being unique. When compared to the automatically created list, the recall was 79%, meaning that 21% of terms extracted manually by evaluators were not extracted by the proposed methodology. All of them had one or more words that were not included in morphological e-dictionaries of Serbian, and therefore no pattern had recognized them. In most cases, those were words specific to the agricultural engineering domain, and not commonly used, such as *samohodna šasija* 'self-propelled chassis'. In some other cases the terms deviated slightly from the used patterns. For example, the term *teško zemljište* is modified within the expression *teško, a plodno, zemljište* 'heavy, but fertile, soil'.

4.1 Comparing the results with the terms extracted by using common association measures

Combination of using syntactic patterns with simple frequency count gave very good results when compared to other approaches. For evaluation purposes, we calculated four more association measures for our term candidates, T-score, the log-likelihood ratio (LLR), C-Value, and Keyness. The term candidates were sorted based on each of the association

measures used and the lists obtained in that way were mutually compared. The excerpt of the first 10 candidates from each list is given in Table X.

Table X *The 10 most frequent candidates, sorted by simple frequency, T-Score, C-Value, LLR and Keyness.*

Simple frequency	**T-Score**	**C-Value**	**LLR**	**Keyness**
električna energija *'electricity'*	električna energija *'electricity'*	električna energija *'electricity'*	brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*
brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*	potrošnja energije *'energy consumption'*	energetska efikasnost *'energy efficiency'*
poljoprivredna proizvodnja *'agricultural production'*	poljoprivredna proizvodnja *'agricultural production'*	poljoprivredna proizvodnja *'agricultural production'*	energetska efikasnost *'energy efficiency'*	potrošnja energije *'energy consumption'*
obrada zemljišta *'soil tillage'*	obrada zemljišta *'soil tillage'*	druga vrsta inertne materije *'other type of inert material'*	poljoprivredna proizvodnja *'agricultural production'*	tehničko rešenje *'technical solution'*
radna brzina *'working speed'*	radna brzina *'working speed'*	obrada zemljišta *'soil tillage'*	električna energija *'electricity'*	vazdušna struja *'windflaw'*
potrošnja energije *'energy consumption'*	potrošnja energije *'energy consumption'*	radna brzina *'working speed'*	obrada zemljišta *'soil tillage'*	brzina vetra *'wind speed'*
energetska efikasnost *'energy efficiency'*	energetska efikasnost *'energy efficiency'*	potrošnja energije *'energy consumption'*	radna brzina *'working speed'*	energija vetra *'wind energy'*
poljoprivredna mehanizacija *'agricultural mechanization'*	poljoprivredna mehanizacija *'agricultural mechanization'*	energetska efikasnost *'energy efficiency'*	poljoprivredna mehanizacija *'agricultural mechanization'*	jedinica površine *'surface unit'*
snaga motora *'engine power'*	snaga motora *'engine power'*	korov druge vrste *'wheed of other type'*	vazdušna struja *'windflaw'*	tehnološki postupak *'technological procedure'*
sistem mašina *'machine system'*	sistem mašina *'machine system'*	poljoprivredna mehanizacija *'agricultural mechanization'*	izvor energije *'source of energy'*	tehnička karakteristika *'technological characteristics'*

The analysis of results showed that simple frequency and T-Score gave comparable results – the first 50 candidates were the same, with slightly different positions (2-3 positions up or down). C-value favorized some longer phrases which are not terms from the domain, but rather some strange word combinations, not so common in language (such as *druga vrsta inertne materije* 'other type of inert matter', which took fourth place in the list). LLR extracted terms that are characteristic for this particular corpus. Because it scored low word combinations common in a language, it is a good option for additional filtering of results. Keyness poorly ranked some very important concepts (highly ranked with other measures).

Precision of term extraction for agricultural and technical terms, for each association measure used in this research is shown in Figures 5 and 6.

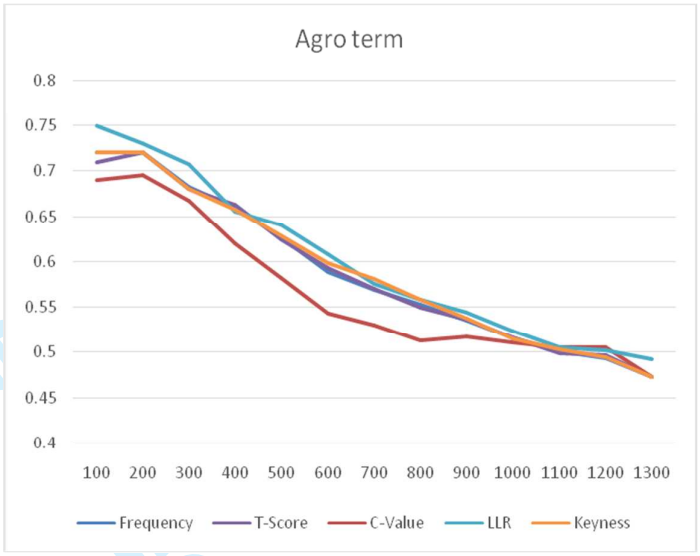

Figure 5. Precision of terms evaluated as agricultural for different association measures

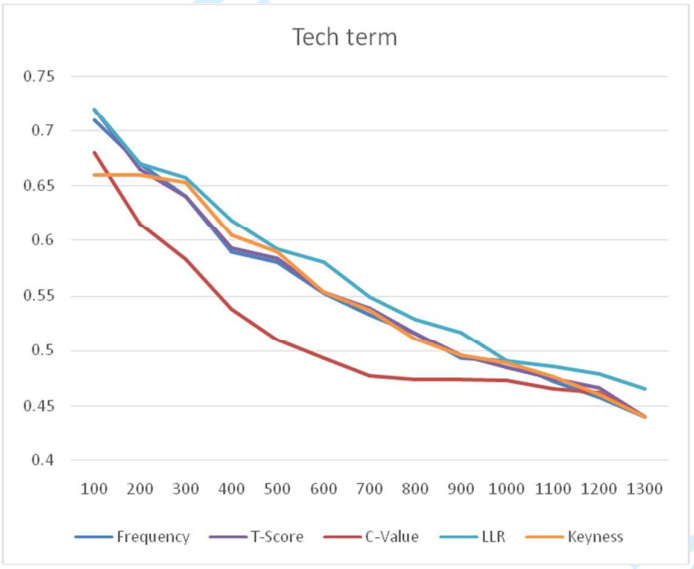

Figure 6. Precision of terms evaluated as technical for different association measures

Mean average precision (MAP) for retrieval of agricultural terms, on all 1523 evaluated terms, ranged from 0.565 (for C-Value) to 0.599 (for LLR). MAP for technical terms ranged from 0.514 (for C-Value) to 0.566 (for LLR). MAP for simple frequency was 0.588 (agricultural terms) and 0.550 (technical terms), calculated for all evaluated terms.

From above comparisons it can be concluded that LLR was the best association measure for term extraction for this agriculture text collection in Serbian. However, the simple frequency gives very similar results as LLR, and can be used as a simpler substitute, in cases when calculating LLR is time-consuming task.

4.2 Extracting terms from domains other than agricultural engineering

The same method for extracting MWTs was applied to texts in Serbian from the mining domain, which was created in a similar research (Stankovic *et al*. (2016)). It contains 10 textbooks, 2 projects and 51 journal articles (with 32,633 sentences and 625,105 simple word forms). The authors used it to automatically both extract multiword expressions from the text and to produce correct lemmas for them, in order to enrich Serbian e-dictionary. The extracted units were not evaluated for being terms from the domain, but only for correctness of the lemma produced. The authors showed that all measures used for precision gave comparable results, with C-value having the highest MAP of 0.804 (the simple frequency had MAP of 0.794). The analysis of extracted MWEs showed that structures AN and NNgi give highest number of MWEs, and are very suitable for automatic lemma production (552 correct lemmas out of 553 extracted with AN, and 599 correct lemmas out of 608 extracted with NNgi).

5 Conclusion and future work

The acquisition of MWTs from the agricultural engineering domain, described in this paper, revealed 870 MWTs not already included in the DELAC dictionary of Serbian. Besides dictionary expansion, this research shows that most MWTs in Serbian have the syntactic structure AN and NNgi. These structures can be used for automatic extraction of terms, but with different levels of precision, with the FST for AN being more precise than the FST for NNgi. After substantially populating our e-dictionaries with terms from the agricultural

domain, FSTs for analyzed structures will retrieve also nested MWTs. For instance, by allowing components to be MWTs themselves the FST that recognizes structures NNgi would retrieve *upravljanje poljoprivrednim zemljištem* 'management of agricultural land', where *upravljanje* 'management' is the first component and *poljoprivrednim zemljištem* 'of agricultural land' is the second component (already in the dictionary of compounds) in the genitive case. In the future, we also plan to test how suitable verb structures (used for some actions) are for automatic extraction. Furthermore, we will try to improve recall by extending patterns so they can recognize some more flexible expressions, as discussed in Section 4.

The approach and methodology used in this research are domain- and language independent. If used for other languages, syntactic structures only need to be adapted to suit the specifics of that particular language.

This research is pioneering work in the natural language processing of agricultural texts in Serbian and, because of that, the results presented in this paper provide resources and tools that are a good foundation for further research and improvements. We plan to continue collecting and processing texts in Serbian, in order to develop an extensive and comprehensive terminological lexicon, not just from the agricultural domain. Furthermore, this kind of text processing, together with language resources, can be used for searching, query expanding, extracting definitions, creation of semantic lexicons, multilingual dictionaries and so on.

Acknowledgements This paper is part of the research funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Ref. No. 178006 and III 47003.

References

Baldwin, T. and Kim, S.N. (2010), "Multiword Expressions", in Indurkhya, N. and Damerau, F.J. (Eds.), *Handbook of Natural Language Processing*, *second edition.* Chapman & Hall/CRC Taylor & Francis Group, FL, pp. 267-292.

Church, K., Gale, W., Hanks, P. and Kindle, D. (1991), "Using Statistics in Lexical Analysis", in: Zernik, U. (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 115-164.

Courtois, B., Silberztein M., et al. (1990), "Dictionnaires électroniques du français". *Langue française*, Vol. 87, pp. 3-4.

Cram, D. and Daille, B. (2016), "Termsuite: Terminology extraction with term variant detection". *Proceedings of the 54th Annual Meeting of the ACL,* pp 13 - 18.

Czerepowicka, M. and Savary, A. (2015), "SEJF - a Grammatical Lexicon of Polish Multi-Word Expressions". *Proceedings of the 7th Language & Technology Conference*, pp 5.

da Graça Krieger, M. and Finatto, M.J.B. (2004), *Introdução à terminologia: teoria e prática.* Contexto, São Paulo.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics*, Vol. 19, pp. 61-74.

Frantzi, K.T., Ananiadou, S. and Tsujii, J. (1998), "The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms". *Research and Advanced Technology for Digital Libraries*, Springer-Verlag, Berlin Heidelberg, pp. 585-604.

Koeva, S. (2007), "Multi-Word Term Extraction for Bulgarian", in *Proceedings of the Workshop on Balto-Slavonic Natural Language: Information Extraction and Enabling Technologies*, Association for Computational Linguistics, PA, USA, pp. 59-66.

Krstev, C., Obradović, I., Stanković, R. and Vitas, D. (2013), "An Approach to Efficient Processing of Multi-Word Units". *Computational Linguistics*, Springer-Verlag, Berlin Heidelberg, pp. 109-129.

Krstev, C., Vitas, D. and Trtovac, A. (2014), "Orwell's 1984 – From Simple to Multi-Word Units". *Human Language Technology Challenges for Computer Science and Linguistics,* Berlin Heidelberg: Springer-Verlag, pp. 276-287.

Liang, Y., Hong, L. and Liu, Y. (2017a), "The Chinese Multi-Word Expression Extraction Based on Improved Semi-supervised Algorithm", in *Proceedings of the 9th International Conference on Machine Learning and Computing* (ICMLC 2017). ACM, New York, NY, USA, 190-194.

Liang, Y,. Tan, H., Li, H., Wang, Z. and Gui, W. (2017b), "A language-independent hybrid approach for multi-word expression extraction", in *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, pp. 3273-3279.

Macken, L., and Tezcan, A. (2016), "Dutch Compound Splitting for Bilingual Terminology Extraction", in: R. Mitkov, J. Monti, G. Corpas Pastor, & V. Seretan (Eds.), *Multi-word Units in Machine Translation and Translation Technology,* John Benjamins.

Mandravickaite, J. and Krilavičius, T. (2017), "Identification of Multiword Expressions for Latvian and Lithuanian: Hybrid Approach". *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, Association for Computational Linguistics, pp.97-101.

Mariani, J. (2005), "Developing Language Technologies with the Support of Language Resources and Evaluation Programs", *Language Resources and Evaluation*, Vol. 39, pp. 35-44.

Nakov, P. and Hearst, M. (2013), "Semantic Interpretation of Noun Compounds Using Verbal and Other Paraphrases". *ACM Transactions on Speech and Language Processing* (TSLP) 2013; 10, 3:13:1-13:51.

Obradović, I. and Stanković, R. (2008), "Software Tools for Serbian Lexical Resources". *INFOtheca – Journal of Informatics & Librarianship*, Vol. 9 No. 1/2, pp. 43a-57a.

Pajić, V., Vujičić Stanković, S. and Pajić, M. (2012), "Transducers for Annotating Weather Information in Meteorological Texts in Serbian". *INFOtheca – Journal of Informatics & Librarianship*, Vol.13 No.2, pp. 36 - 51.

Przepiórkowski, A., Degórski, Ł., Wójtowicz, B., Spousta, M., Kuboň, V., Simov, K., Osenova, P. and Lemnitzer, L. (2007), "Towards the automatic extraction of definitions in Slavic". *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, Association for Computational Linguistics, PA, USA, pp. 43-50.

Ramisch, C. (2015), *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing, Springer International Publishing, ISBN 978-3-319-09206-5.

Romary, L. (2014), "TBX goes TEI – Implementing a TBX basic extension for the Text Encoding Initiative guidelines". arXiv preprint arXiv:1403.0052.

Savary, A. (2008), "Computational Inflection of Multi-Word Units – A contrastive study of Lexical Approaches". *Linguistic Issues in Language Technology*, Vol. 1 No. 2, pp.1-53.

Savary, A. and Zaborowski, B. (2012), "SEJFEK – a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units". *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon* (*CogALex-III*), Mumbai, India, pp. 195-214.

Sclano, F. and Velardi, P. (2007), "TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities". *Enterprise Interoperability II*, Springer, London, pp. 287-290.

Scott, M. and Tribble, C. (2006), *Textual Patterns: Keyword and Corpus Analysis in Language Education*, John Benjamins, Philadelphia.

Stankovic, R., Krstev, C., Obradovic, I., Lazic, B. and Trtovac, A. (2016), "Rule-based Automatic Multi-word Term Extraction and Lemmatization", in: Calzolari, N., et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC 2016*), European Language Resources Association (ELRA), Paris, France, pp. 507-514.

Stanković, R., Obradović, I., Krstev, C. and Vitas, D. (2011), "Production of morphological dictionaries of multi-word units using a multipurpose tool", in: Jassem, K., Fuglewicz, P., Piasecki, M. and Przepiorkowski, A. (Eds.), *Proceedings of the Computational Linguistics-Applications Conference*, Jachranka, Poland, pp. 77-84.

Tadić, M. and Šojat, K. (2003), "Finding Multiword Term Candidates in Croatian". *Proceedings of Information Extraction for Slavic Languages 2003 Workshop*, BAS, Sofija, pp. 102-107.

Tsvetkov, Y. and Wintner, S. (2014), "Identification of multiword expressions by combining multiple linguistic information sources". *Computational Linguistics*, 40(2):449–468.

Utvić, M. (2011), "Annotating the Corpus of contemporary Serbian", *INFOtheca – Journal of Informatics & Librarianship*, Vol. 12 No. 2, pp. 36a-47a.

Verberne, S., Sappelli, M., Hiemstra, D. and Kraaij, W. (2016), "Evaluation and analysis of term scoring methods for term extraction." *Information Retrieval Journal*, Volume 19, Issue 5, pp 510–545.

Viera, A.J., Garrett, J.M., et al. (2005), "Understanding Interobserver Agreement: The Kappa Statistic". *Family Medicine*, Vol. 37, pp. 360-363.

Vintar, Š. (2004), "Comparative Evaluation of C-value in the Treatment of Nested Terms", in *Proceedings of the Language Resources and Evaluation Conference MEMURA 2004 Workshop* (*Methodologies and Evaluation of Multiword Units in Real-world Applications*), Lisbon, Portugal, pp. 54-57.

Vitas, D., Popović, Lj., Krstev, C., Obradović, I., Pavlović-Lažetić, G. and Stanojević, M. (2005), *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age. META-NET White Paper Series*, Springer-Verlag, Berlin Heidelberg.

Vivaldi, J. and Rodríguez, H. (2007), "Evaluation of terms and term extraction systems: A practical approach". *Terminology*, Vol. 13 No. 2, pp. 225-248.

Yin, Y., Wei, F., Dong, L., Xu, K., Ming, Z. and Ming, Z. (2016), "Unsupervised Word and Dependency Path Embeddings for Aspect Term Extraction". *CoRR Computation and Language*, http://arxiv.org/abs/1605.07843.

Zaninello A. and Nissim M. (2010), "Creation of Lexical Resources for a Characterisation of Multiword Expressions in Italian", in: Calzolari N. et al., editors. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* pp. 654-661.

Zhang, C. and Wu, D. (2012), "Bilingual terminology extraction using multi-level termhood". *The Electronic Library*, Vol. 30 No.2, pp. 295-309.

Semi-Automatic Extraction of Multiword Terms from Domain-Specific Corpora

TABLES

Table I Syntactic patterns used for MWT extraction.

ID	Syntactic pattern*	Description	Example
1	AN	Both components inflect and agree in gender, number and case	*žitni kombajn* 'wheat harvester'
2	X-N	The first component does not inflect (and is never used as an independent word in Serbian); the second component inflects	*eko-sistem* 'ecosystem'
3	NNgi	Only the first component inflects; the second is always in genitive	*obrada zemljišta* 'soil tillage', lit. 'tillage of soil'
4a	NPrepNp	Only the first component inflects, followed by prepositional phrase	*sistem za navodnjavanje* 'irrigation system', lit. 'system for irrigation'
4b	NNgiNgi	The first component inflects; the second and third are always in genitive	*tok vegetacionog perioda* 'course of vegetation period'
5	ANNgi	The first and the second components inflect, the third is always in genitive	*unutrašnji deo parcele* 'inner part of a parcel'
6	N(-)N	Both components inflect and agree in case and number; can be optionally separated by a hyphen	*traktor guseničar* 'tractor caterpillar'
7	AAN	All components inflect and agree in gender, number and case	*elektronska komandna jedinica* 'electronic command unit'
8a	NNgiPrepNp	Only the first components inflects; the second is always in genitive, followed by prepositional phrase	*sistem mašina za doradu* 'machine system for final treatment'
8b	NNgiNgiNgi	Only the first components inflects; the rest are always in genitive	*redukcija stepena zagađenja vazduha* lit. 'reduction of the level of the pollution of air'
8c	NPrepNpNgi	The first component inflects;it is followed by prepositional phrase, the fourth is in genitive	*mašina za obradu zemljišta* lit. 'machine for tillage of soil'
9a	ANPrepNp	The first two components inflects and agree in case, gender and number; the last component is prepositional phrase	*tehnički sistemi za redukciju* 'technical systems for reduction'
9b	ANNgiNgi	The first two components inflects and agree in case, gender and number; the last two components are always in genitive	*energetska upotreba drvne biomase* lit. 'energetic utilization of tree biomass'

10	X-AN	The first component does not inflect (and is never used as an independent word in Serbian); the second and third component inflect and agree in case, number and gender	*tehničko-tehnološko rešenje* lit. 'technical-technological solution'

* A – adjective, N – noun, Ngi – noun or adjective in the genitive or instrumental case, "-" – the separating hyphen, (-) – optional occurrence of the separating hyphen, PrepNp – prepositional phrase, X – invariable morpheme.

Table II *Extracted term candidates, with their normalized forms and graph labels.*

Term candidate occurrence	Normalized form (lemmatized word by word)	Graph pattern	English translation
agrarne politike agrarnom politikom agrarnu politiku	agrarni politika	AN	agricultural policy
merač pritiska	merač pritiska	NNgi	pressure gauge
merača pritiska	merač pritiska	NNgi	pressure gauge

Table III *An excerpt from the table with numbers of occurrence of term candidates, grouped by normalized form and graph. It can be seen that the two occurrences were recognized by using two different graphs, producing different normalized forms.*

Graph	Number	Word form	Temporary lemma	Lemma	Frequency
NNgi	plu	rezultata istraživanja	rezultat istraživanja	rezultati istraživanja	31
		rezultate istraživanja			2
		rezultati istraživanja			53
		rezultatima istraživanja			30
	sin	rezultat istraživanja		rezultat istraživanja	6
		rezultata istraživanja			31
		rezultate istraživanja			2
		rezultatu istraživanja			1
N-N	sin	rezultata istraživanja	rezultat istraživanje	rezultat istraživanje	31

Table IV *An excerpt from the list of terms with semantic markers.*

MWT	Semantic markers	English translation
brzina kretanja	+DOM=*Tech*	moving speed
izvor energije	+DOM=*Agr*+DOM=*Tech*	energy source
poljoprivredna proizvodnja	+DOM=*Agr*	agricultural production

Table V *The results of extraction of term candidates after the STEPS 3 - 7.*

Pattern	Term candidates	Evaluated as MWT	Correct lemma
AAN	1	1	1
AN	417	406	386
ANNgi	74	59	67
ANNgiNgi	13	8	12
ANPrepNp	20	12	14

N-N	158	66	10
NNgi	526	459	476
NNgiNgi	130	106	114
NNgiNgiNgi	9	7	9
NNgiPrepNp	11	9	11
NprepNp	137	87	109
NPrepNpNgi	20	17	17
X-AN	6	4	3
X-N	1	1	1
Total	**1523**	**1242**	**1230**

Table VI *An excerpt from the resulting table with the most frequent terms and their evaluation.*

Term candidate	**Evaluator 1**		**Evaluator 2**		**Frequency**	**IT**	**P@n**
	Agr	**Tech**	**Agr**	**Tech**			
električna energija 'electricity'	no	yes	no	yes	346	yes	1.00
brzina kretanja 'moving speed'	no	yes	no	yes	320	yes	1.00
poljoprivredna proizvodnja 'agricultural production'	yes	no	yes	no	211	yes	1.00
obrada zemljišta 'soil tillage'	yes	yes	yes	yes	200	yes	1.00
radna brzina 'working speed'	no	yes	no	yes	196	yes	1.00
potrošnja energije 'energy consumption'	no	yes	no	yes	195	yes	1.00
energetska efikasnost 'energetical efficiency'	no	yes	no	yes	180	yes	1.00
poljoprivredna mehanizacija 'agricultural mechanization'	yes	yes	yes	yes	146	yes	1.00
snaga motora 'power of engine'	no	yes	no	yes	143	yes	1.00
sistem mašina 'machine system'	no	yes	no	yes	133	yes	1.00

Table VII *Evaluator's agreement.*

	Agr			**Tech**		
	Eval(1)	**Eval(2)**	**Agree**	**Eval(1)**	**Eval(2)**	**Agree**
ny*	1053	1076	486	1103	1094	461
nn	587	564	975	537	546	1018
$Total$	1640	1640	1461	1640	1640	1479
$p(y)$	0.642	0.656		0.673	0.667	
$p(n)$	0.358	0.344		0.327	0.333	
$Pr(a)$			0.891			0.902
$Pr(e)$			0.544			0.558
κ			0.760			0.778

* ***ny*** – number of 'yes' answers, ***nn*** – number of 'no' answers.
Total – total number of answers (total number of answers where both evaluators agree).
p(y) – probability of a 'yes' answer, $p(n)$ – probability of a 'no' answer.
Pr(a) – probability of an answer where both evaluators agree,
Pr(e) = $p_1(y) * p_2(y) + p_1(n) * p_2(n)$, $\boldsymbol{\kappa} = (Pr(a) - Pr(e))/1 - Pr(e)$.

Table VIII *The frequencies and precision for first n term candidates.*

n	*freq(n)*	*prec(n) = P@n*	*n*	*freq(n)*	*prec(n) = P@n*
1	346	1	800	14	0.68
100	48	0.87	900	12	0.66
200	33	0.87	1,000	12	0.65
300	26	0.86	1,100	11	0.63
400	21	0.81	1,200	10	0.62
500	18	0.77	1,300	10	0.61
600	16	0.74	1,400	10	0.60
700	15	0.71	1,500	9	0.58

Table IX *The number of terms extracted by particular graphs, i.e. having particular structures (out of 1,744 evaluated terms).*

Syntactic structure	**Total number of term candidates (n_{tc})**	**Total number of terms (n_t)**	**n_t/n_{tc} ratio**
AAN	1	1	1.00
AN	469	380	0.81
ANNgi	72	39	0.54
ANNgiNgi	9	1	0.11
ANPrepNp	22	5	0.23
N-N	181	70	0.39
NNgi	493	300	0.61
NNgiNgi	115	51	0.44
NNgiNgiNgi	7	1	0.14
NNgiPrepNp	10	2	0.20
NprepNp	120	29	0.24
NPrepNpNgi	16	3	0.19
X-AN	7	5	0.71
X-N	1	0	0.00

Table X *The 10 most frequent candidates, sorted by simple frequency, T-Score, C-Value, LLR and Keyness.*

Simple frequency	**T-Score**	**C-Value**	**LLR**	**Keyness**
električna energija *'electricity'*	električna energija *'electricity'*	električna energija *'electricity'*	brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*
brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*	brzina kretanja *'moving speed'*	potrošnja energije *'energy consumption'*	energetska efikasnost *'energy efficiency'*
poljoprivredna proizvodnja *'agricultural production'*	poljoprivredna proizvodnja *'agricultural production'*	poljoprivredna proizvodnja *'agricultural production'*	energetska efikasnost *'energy efficiency'*	potrošnja energije *'energy consumption'*
obrada zemljišta *'soil tillage'*	obrada zemljišta *'soil tillage'*	druga vrsta inertne materije *'other type of inert material'*	poljoprivredna proizvodnja *'agricultural production'*	tehničko rešenje *'technical solution'*
radna brzina *'working speed'*	radna brzina *'working speed'*	obrada zemljišta *'soil tillage'*	električna energija *'electricity'*	vazdušna struja *'windflaw'*
potrošnja energije *'energy consumption'*	potrošnja energije *'energy consumption'*	radna brzina *'working speed'*	obrada zemljišta *'soil tillage'*	brzina vetra *'wind speed'*
energetska efikasnost *'energy efficiency'*	energetska efikasnost *'energy efficiency'*	potrošnja energije *'energy consumption'*	radna brzina *'working speed'*	energija vetra *'wind energy'*
poljoprivredna mehanizacija *'agricultural*	poljoprivredna mehanizacija *'agricultural*	energetska efikasnost *'energy efficiency'*	poljoprivredna mehanizacija *'agricultural*	jedinica površine *'surface unit'*

mechanization'	*mechanization'*		*mechanization'*	
snaga motora *'engine power'*	snaga motora *'engine power'*	korov druge vrste *'wheed of other type'*	vazdušna struja *'windflaw'*	tehnološki postupak *'technological procedure'*
sistem mašina *'machine system'*	sistem mašina *'machine system'*	poljoprivredna mehanizacija *'agricultural mechanization'*	izvor energije *'source of energy'*	tehnička karakteristika *'technological characteristics'*