Terminology Acquisition and Description Using Lexical Resources and Local Grammars

Cvetana Krstev, Ranka Stanković, Ivan Obradović, Biljana Lazić

Terminology acquisition and description using lexical resources and local grammars

Cvetana Krstev
University of Belgrade
cvetana@matf.bg.ac.rs

Ranka Stanković
University of Belgrade
ranka@rgf.bg.ac.rs

Ivan Obradović
University of Belgrade
ivan.obradovic@rgf.bg.ac.rs

Biljana Lazić
University of Belgrade
biljana.lazic@rgf.bg.ac.rs

Abstract

Acquisition of new terminology from specific domains and its adequate description within terminological dictionaries is a complex task, especially for languages that are morphologically complex such as Serbian. In this paper we present an approach to solving this task semi-automatically on basis of lexical resources and local grammars developed for Serbian. Special attention is given to automatic inflectional class prediction for simple adjectives and nouns and the use of syntactic graphs for extraction of Multi-Word Unit (MWU) candidates for termbases, their lemmatization and assignment of inflectional classes.

1 Introduction

In this paper we present a semi-automatic procedure for terminology acquisition in Serbian. Rapid changes in many knowledge domains mean that new terms are continuously being created and introduced in Serbian making important the automation of their retrieval and incorporation in Serbian terminological dictionaries. Due to specific features of Serbian grammar, especially its rich morphology, this is a complex task, and corresponding language resources in the form of morphological e-dictionaries and grammars need to be applied (Vitas et al., 2012). For that reason, in the case of Serbian, it is not enough to extract terminology from the domain, but it also has to be adequately described, for instance, in the form of e-dictionaries.

The field of terminology is strongly related to research on multiword terms, which relates closely to MWEs (Baldwin & Kim, 2010; Frantzi et al., 2000). An analysis of terms from technical dictionaries for different domains (fiber optics, medicine, physics and mathematics, psychology) showed that 97% of multi-words in these sources consist of nouns and adjectives only, and more than 99% consist only of nouns, adjectives, and a preposition. (Justeson & Katz, 1995) Identifying the adjectives and the prepositional phrase is thus important for terminology acquisition (Daille, 2000).

There are two mainstream approaches (Enguehard & Pantera, 1995; Cerbah & Daille, 2007) to terminology acquisition. One relies on using statistical measures (Nakagawa & Mori, 2003; Ramisch et al., 2012; Quochi et al., 2012; Zhang et al., 2006) and the other is based on linguistic rules. A rule-based approach for the extraction of terms based on a cascade of transducers using CasSys tool incorporated in Unitex[1] corpus processing platform, as well as the use of TMF standard for the representation of terms is proposed in (Ammar et al., 2015) and applied on Arabic scientific and technical corpus. In (Savary et al., 2012) terminology extraction in the domain of ecsonomy is presented for Polish. It has two modules: a grammatical lexicon of terminological MWEs and a fully lexicalized shallow grammar, obtained by an automatic conversion of the lexicon. Przepiorkowski and associates (2007) present results of automatic extraction of term definitions from unstructured texts in Bulgarian, Czech and Polish by use of regular grammars.

There are also combinations of the two approaches (Rodrıguez et al., 2007). Sag et al. reported that modern statistical Natural Language Processing (NLP) is in great need of better language models and linguistic tools must come to

[1] Corpus processing System Unitex: http://www-igm.univ-mlv.fr/~unitex/

grip with problems of disambiguation and MWUs (Sag et al., 2002).

2 Process description

The processing steps (Fig.1) of integrating new terms from specific domains in terminological dictionaries using lexical resources and local grammars in our approach are:

1. Linguistic preprocessing of the input plain text file from the chosen domain using Unitex.
2. Analysis of unrecognized words as the most probable source of terminology and expanding the dictionary of simple words:
   - 2.1 Retrieval of unrecognized words;
   - 2.2 Manual filtering, preparation of a list of extracted terms in canonical forms (for instance, nominative singular for nouns) and annotating with semantic labels (e.g. human) and some grammatical categories (e.g. adding the gender for the nouns);
   - 2.3 Automatic prediction of the inflectional class and the production of dictionary entry in DELA format (detailed description of the algorithm is given in section 3);
   - 2.4 Compiling the dictionaries of newly acquired terms and integrating them with other resources for linguistic text processing;
   - 2.5 Repeated linguistic preprocessing with expanded dictionaries for verification of recognition of new lemmas.
3. MWUs extraction
   - 3.1. Application of syntactic graphs to extract MWUs with different syntactic structures from the same text (detailed description of the algorithm is given in section 4);
   - 3.2. Removing duplicate extractions: if a sequence of words is recognized with different graphs as having different syntactic structures the most probable candidate is chosen according to the pre-established order of precedence;
   - 3.3. Two-step generation of MWU canonical forms: in the first step lemmatization of simple words that form the MWU is performed, while in the second step the lemma of the MWU is produced based on the results from step 1.
4. Selection of terms from new MWUs
   - 4.1. Frequency calculation for all forms of MWUs and their basic forms with ranking of results;
   - 4.2. Removing MWUs already in e-dictionaries and those with rank under the specified threshold;
   - 4.3. Linguistic evaluation of grammatical correctness of remaining MWUs;
   - 4.4. Assessment of domain relevance of each MWU by comparing its frequency in the domain text with its frequency in the Corpus of Contemporary Serbian (Utvić, 2014).
5. Expanding MWU dictionaries
   - 5.1. Creation of complete MWU lemmas in compliance with DELAC format (Savary, 2009);
   - 5.2. Compiling the dictionaries of newly acquired multi-word terms and integrating them with other resources for linguistic text processing;
   - 5.3. Linguistic pre-processing with expanded dictionaries for verification of recognition of new MWU lemmas.

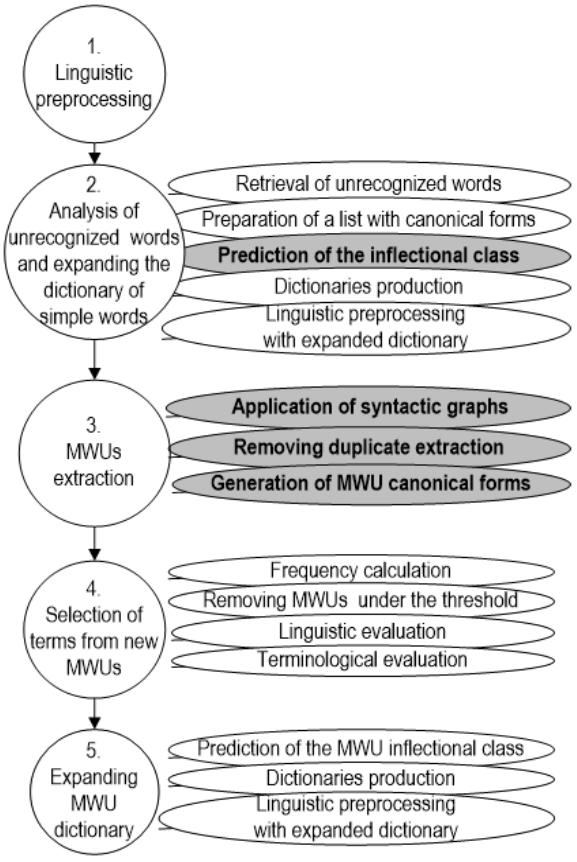

Figure 1: Diagram of terminology acquisition using lexical resources and local grammars

The newly acquired terms, both simple and MWU, can be exported to termbases, TBX and other standard formats for terminological resources. In this paper we will focus on (marked

gray in Figure 1): inflectional class prediction (step 2.3) and extraction of MWU candidates for termbases using syntactic graphs (step 3).

3 Prediction of inflectional class for simple words

Prediction of inflectional class for a new word in Serbian is not an easy task because of complex inflectional grammar with numerous rules and exceptions. Morphological electronic dictionaries of Serbian for NLP are being developed for many years now. Their development follows the methodology and format (known as DELAS/DELAF) presented for French in (Courtois, 1990). E-dictionaries in the same format have been produced for many other languages.

In dictionary of lemmas (DELAS) each lemma is described in full detail so that a dictionary of forms containing all necessary grammatical information (DELAF) can be generated from it, and subsequently used in various NLP tasks.

Serbian e-dictionaries of simple forms have reached a considerable size: they have about 135,000 lemmas generating more than 5 million forms and 13,000 compound lemmas, that is, multi-word units (Krstev, 2008). The number of simple lemmas by Part-Of-Speech (POS) is depicted in Figure 2 (left).

POS	lemmas		FSTs	
Nouns	81,866	61%	372	44%
Verbs	17,071	13%	372	44%
Adjectives	31,071	23%	69	8%
Other	4,632	3%	41	5%
Total	134,640		854	

Figure 2: Statistics of lemmas and inflectional FSTs

Inflectional classes are described with metadata including most important features for class distinction e.g. for nouns grammatical gender and number, case, and animateness are given.

Grammatical inflectional rules are encoded by 854 inflectional Finite-State Transducers (FST) Inflectional FSTs are a special kind of FSTs used for modeling inflectional paradigms, that is, inflectional classes. Each FST of this kind is used for production of all inflected forms for all lemmas belonging to the same class. The number of Inflectional FSTs by POS is depicted in Figure 2 (right).

Productiveness of all inflectional classes are not the same: some classes are used for a large number of regular cases, while other pertain to (rare) exceptions. Our approach is addressing the first group, having in mind that terminology usually inflects regularly. Figure 3 presents the number of inflectional classes and percent of lemmas that belong to them. For example, 10 classes for adjectives account for 98% of lemmas, 10 classes for nouns account for 61.8% of lemmas, and 10 classes for verbs account for 59.6% of lemmas.

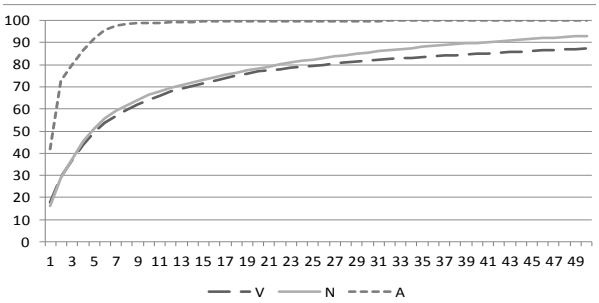

Figure 3: FST classes and the percentage covering the dictionary of lemmas

FST class prediction can be divided into two parts: one is extraction of implicit knowledge and the other is actual prediction of FST class for a new lemma. Extraction of implicit knowledge in the form of a dataset with word endings, grammatical categories and FST classes proceeds as follows:

1. Calculate frequencies for each POS and relative frequencies for each FST class within POS in the current dictionary of simple lemmas.
2. Create a dataset from DELAS lemma endings of length 3,4,5 and 6 characters with corresponding grammatical categories retrieved from DELAF (e.g. for nouns in that dataset: POS, lemma, FST, gender, animateness, pronunciation).
3. Create another dataset with frequencies for each combination of FST code and grammatical category and for each ending of length 3,4,5,6, as an estimate of the probability that the FST class is the appropriate one. The dataset includes: ending, POS, gender, animateness, pronunciation, FST and probability (chance rank 0-100) for FST (table 1, column Rel. freq.).

ending	POS	length	gender	anim	FST	Total	Frequency	Rel. freq.	Example
alica	N	5	f	-	N650	97	95	98	sij**alica**
alica	N	5	f	-	N1650	97	1	1	Sk**alica**
alica	N	5	f	+	N651	27	26	96	var**alica**
alica	N	5	f	+	N1651	27	1	4	L**alica**
ica	N	3	m	+	N1683	145	142	98	Milo**jica**
ica	N	3	m	+	N1741	145	3	2	Pr**ica**
ica	N	3	m/f	+	N683	40	33	83	tvrd**ica**
ica	N	3	m/f	+	N679	40	7	18	ub**ica**

Table 1: Excerption from dataset with ending.
(m: masculine gender; f: feminine gender; m/f: nouns change gender in their inflectional paradigm)

Analysis of the relation between word endings and inflectional FST classes shows that the prediction of inflectional class by the abovementioned statistical analysis of existing dictionaries is justified. Figure 4 illustrates this relation for word endings of length 3, 4, 5 and 6. For example, in the case of word endings of length 3, for 33% of words from the existing dictionary there is only one corresponding FST class, for approximately 20% of words there are two classes, and so on, whereas for word endings of length 6 there is a single class for as much as 90% of words.

In order to facilitate prediction of FST class, a set of rules based on inflectional class metadata is used. Distinction between inflectional classes based on grammatical categories can be done only to some extent, so implicit knowledge from the existing dictionary of simple words is used to improve prediction.

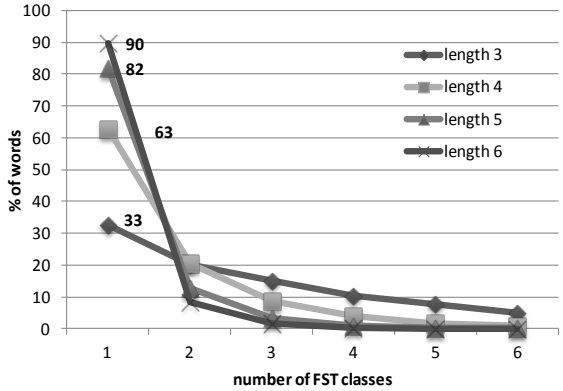

Figure 4: Relation between word endings and inflectional FST classes

The process of automatic prediction of inflectional FST class for a new entry follows a hybrid approach: one part is rule-driven with explicit codification of knowledge about FST classes and the other is statistical, based on existing dictionary of simple word lemmas with implicit knowledge about dependence between FST classes and dictionary entries.

After preparing the list of new entries in the form: lemma, POS, Grammatical_Categories (e.g. *grabuljar*,*N*,*Rud* 'rake') the following procedure is applied:

1. For each candidate lemma filter the dataset prepared from previous step as follows:
   1.1. if the lemma has specific marks for pronunciation, then retain only dataset members with the same mark and remove the rest;
   1.2. if the grammatical gender or animateness is assigned, retain only dataset members with the same grammatical category and remove the rest;
   1.3. if the first letter of the lemma is in upper case additional filtering can take place taking into account FST classes which have only inflected singular forms.
2. After filtering and ranking the dataset, prediction (FST assignment) for the lemma is repeated with threshold from 99 to 95 for relative frequency, for suffixes 6,5,4, and 3 respectively;
3. For thresholds under 95 and over 80 lemma prefix (if longer than 2 characters) is used: if the prefix is in the dictionary of prefixes and the remainder of the lemma is a word in DELAS, then the lemma is the inflectional class of the corresponding DELAS word is assigned to the lemma.
4. For thresholds 80 and less steps 1 and 2 only are repeated.

From a sample of domain texts and dictionaries we manually filtered 623 new terms from domains of mining, geology and e-learning and applied the described procedure for FST class prediction: to 582 (93%) of them the correct FST

class was assigned, 27 (4%) had a partly correct class assigned (for instance, inflection is correct but falsely allows plural forms), and to 14 (2%) of them an incorrect class was assigned.

4 Syntactic graphs for MWU recognition

4.1 Structure of terms in termbases

In order to analyze the structure of terms in different domains, primarily the number of components they consist of, we used samples from three terminological resources for Serbian. Two terminological resources, GeolISSTerm [2] and RudOnto[3] have been developed at University of Belgrade, Faculty of Mining and Geology. GeolISSTerm is a bilingual thesaurus of geological terms in Serbian and their English equivalents (Stankovic et al., 2011), divided in several subdomains: petrology, mineralogy, hydrogeology, geophysics, structural geology etc. RudOnto is covering the larger area of mining engineering and mine safety terminology (Stankovic et al., 2012). The third termbase used is the Dictionary of Library and Information Sciences (RNBS),[4] developed by the National Library of Serbia. It contains terminology in Serbian, English and German, related to theory and practice of librarianship and information sciences and a wide range of close or related fields.

Table 2. Frequencies of terms of different lengths in samples from 3 termbases

Dictionary	Term length (in number of words)				
	1	2	3	4	≥5
GeolISS Term	1436	2356	749	305	243
RNBS	3302	6180	2062	806	415
RudOnto	1004	1351	1350	1031	2341

Table 2 presents the distribution of terms consisting of 1, 2, 3, 4 and more components for the three termbases. These results are consistent with the results presented in (Justeson et al., 1995), at least for GeolISSTerm and RNBS, and show that terms with 5 or more components are much less frequent than the shorter ones. The results are somewhat different for RudOnto, as it contains very specific terms, such as causes of injuries, employee positions, types of injuries, or technical characteristics of machines, which are often longer MWUs than the less specific terminology of the two other termbases. Two examples from RudOnto can illustrate this: a term for employee position "Geologist for mineralogy, petrology, sedimentology and geochemical research" and a term for technical characteristics of machines "Length of the caterpillar transporting device measured from the vertical excavator rotation axis to the front edge of the caterpillar".

4.2 Extraction of MWUs from domain texts

The extraction of MWUs from a text is preceded by the retrieval of new simple word terms from it and their incorporation in the existing system of morphological e-dictionaries as MWU extraction relies heavily on existing lexical resources.

In the Serbian e-dictionary of MWUs, all entries are distributed in classes according to their syntactic structure, or more precisely, according to the information needed for their inflection. The names of classes correspond to the names of special FSTs that are used for MWU inflection. For instance, the class AXN pertains to MWUs with the syntactic structure: an adjective (A) followed by a noun (N), where the two components agree in gender, number, case and animateness. In class names X stands for a component that does not inflect when a MWU inflects or for a component separator. In the case of AXN, X stands for the separator, usually a space. Sometimes, MWUs with different syntactic structure belong to the same class. For instance, the class N4X implies that MWUs belonging to it consist of a noun followed by two other components (separated by two separators) that do not inflect. The syntactic structure of these components can be a noun followed by two adjectives/nouns in the genitive case (e.g. *eksploatacija mineralnih sirovina* 'exploitation of mineral resources') but also a noun followed by a prepositional phrase (e.g. *bager na šinama* 'excavator on rails').

There are 29 such classes for Serbian nominal MWUs.[5] However, 10 of them are used for the inflection of more than 98% of all nominal MWUs. Four of these classes are used for the inflection of two component MWUs, four for the inflection of 3-component MWUs and two for the inflection of 4-component MWUs. Given that

[2] http://geoliss.mprrpp.gov.rs/term
[3] http://rudonto.rgf.bg.ac.rs/
[4] http://rbi.nb.rs/en/home.html

[5] The number of FSTs (80) is greater than the number of classes because they deal with other details of inflection: does the MWU inflect in number, are some components optional, etc.

they cover the large majority of MWUs, we have developed syntactic FSTs for the extraction of MWUs belonging to these 10 classes. They are, listed in the descending order of their frequency:

1. **AXN** – an adjective followed by a noun; the adjective and the noun have to agree in all four grammatical categories; e.g. *zemni gas* 'natural gas'.
2. **2XN** – a noun preceded by a word that does not inflect in the MWU. Usually it is a word used only in one or a few MWUs, a prefix or an adverb derived from an adjective, while the separator is usually a hyphen; e.g. *anker-mreža* 'anchor network'.
3. **N2X** – a noun followed by a word that does not inflect in the MWU. Usually this word is a noun in the genitive or in the instrumental case; e.g. *patrona eksploziva* 'explosive cartridge' and *upravljanje krovinom* 'roof control'.
4. **N4X** – a noun followed by two words that do not inflect in the MWU. Two syntactic structures are possible:
   a. **NNgi** - A noun followed by two adjectives/nouns in the genitive case or in the instrumental case; e.g. *otkopavanje širokim čelom* 'broad forehead excavation'.
   b. **NprepNp** - A noun followed by a prepositional phrase; e.g. *lanac sa grabuljama* 'chain with a rake'.
5. **AXN2X** – a noun preceded by an adjective that agrees with it in gender, number, case and animateness and followed by a word that does not inflect in the MWU, usually a noun in the genitive or instrumental case; e.g. *geološko kartiranje terena* 'geological field mapping'.
6. **NXN** – a noun followed by a noun that agrees with it in number and case, where the separator can be a hyphen; e.g. *bager kašikar* 'shovel excavator'.
7. **AXAXN** – a noun preceded by two adjectives that agree with it in gender, number, case and animateness; e.g. *površinski istražni radovi* 'surface exploration works'.
8. **N6X** - a noun followed by three words that do not inflect in the MWU. Three syntactic structures are possible:
   a. **NNgiPrepNp** - a noun followed by a noun in the genitive case and a prepositional phrase (as in case 4b); e.g. *priprema ležišta za otkopavanje* 'deposit preparation for mining'.
   b. **NNgiNgiNgi** - a noun followed by three nouns/adjectives in the genitive case; e.g. *istraživanje ležišta mineralnih sirovina* 'exploration of mineral deposits'.
   c. **NprepNpNgi** - a noun followed by a prepositional phrase; e.g. *bakar sa primesama zlata* 'copper with a sprinkling of gold'.
9. **AXN4X –** a noun preceded by an adjective that agrees with it in gender, number, case and animateness and followed by two words that do not inflect in the MWU. Two syntactic structures are possible:
   a. **ANPrepNp** - A noun preceded by an adjective and followed by a prepositional phrase (as in case 4b); e.g. *gravitacijska koncentracija u vodi* 'gravity concentration in water'.
   b. **ANNgiNgi** - a noun preceded by an adjective and followed by two adjectives/nouns in the genitive case or in the instrumental case (a 4a case); e.g. *površinska eksploatacija mineralnih sirovina* 'surface exploitation of mineral resources'.
10. **2XAXN** - an adjective followed by a noun that agrees in all four grammatical categories and preceded by a word that does not inflect in the MWU; e.g. *magmatsko-eruptivni masiv* 'magmatic-igneous massif'.

FST for extraction of MWUs of type AXN with two paths from one of the subgraphs that illustrate the agreement between adjectives and nouns is depicted in Figure 5. Dictionary variable used for FST output in the form $a.LEMMA$ retrieves a lemma of recognized word form a thus performing the simple word lemmatization.

Due to high homography of word forms it may happen that the same sequence of words is recognized by two or more graphs; naturally, only one recognition may be correct. For instance if the MWU *bager kašikar* (case 6, NXN) is detected in the analyzed text in the genitive case *bagera kašikara* it may be erroneously interpreted as a MWU of a form NNg (case 3) in the genitive case. Consequently, all NNg constructions in an analyzed text that appear in the genitive case (which happens very frequently) will be interpreted also as a NXN case. For that reason, in the case of ambiguous recognition we always give precedence to the more probable case. For instance, for 2-component MWUs the precedence is: AXN, 2XN, N2X, NXN.

As a rule, we are looking for the longest match for a MWU, that is, if a text matches an

AXAXN pattern, than we will ignore the match AXN that is subsumed. However, in certain cases we take into consideration the shorter matches as well. For instance, a sequence recognized as NNgNgNg, may well not be a multi-unit term, but rather consist of two multi-unit terms of the form NNg or contain as its part a AXAXN term; e.g. *sprečavanje zagađenja životne sredine* 'prevention of environmental pollution' may not be considered a term, while *zagađenje životne sredine* '*environmental pollution*' is. For that reason, the order of term candidate extraction is:

1. AXAXN, 2XAXN, AXN2X, AXN4X, AXN
2. N6X
3. N4X
4. 2XN, N2X, NXN

At the end of each round duplicates are eliminated according to the priorityand the union of all results is performed.

The output of processing by transducers is the initial version of the normalized MWU that consists of simple word lemmatization — inflected parts of a MWU are replaced by their lemmas, as they are recorded in e-dictionaries. The list of produced normalized MWUs is then additionally processed by a new set of transducers in order to obtain correct MWU lemmas. The following adjustments have to be performed:

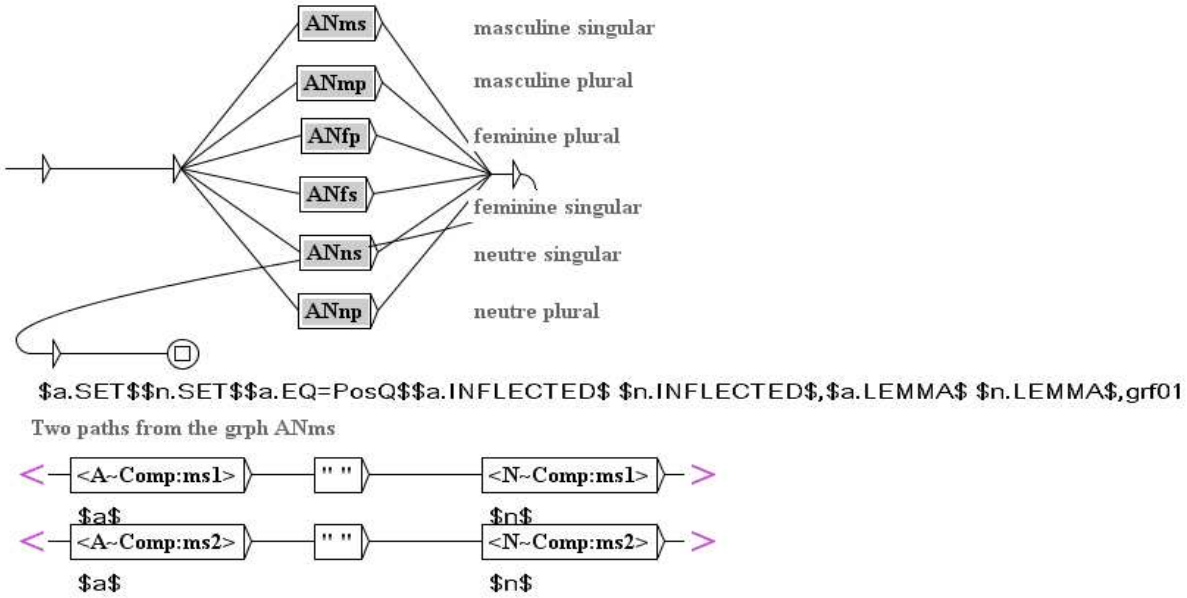

Figure 5. An FST for extraction of MWUs

1. For MWUs with syntactic structure AXN, AXAXN, AXN2X, AXN4X, and 2XAXN the form of the adjectives has to be corrected so that the right gender is selected to correspond to the gender of the noun (simple word lemmas are always in the masculine gender). For example, when simple word lemmatization offers a lemma $minski_m$ $bušotina_f$ 'blasting boreholes' it has to be corrected to $minska_f$ $bušotina_f$.
2. For all MWUs, the right number of the MWU has to be selected: if it appeared in a text only in singular form or only in plural form, then the lemma will be in the respective form (e.g. only singular form *jamski vazduh* 'air in the underground mine', only plural form *atmosferske padavine* 'atmospheric precipitation'); if it appeared in both plural and singular forms, then both forms of lemmas will be offered.

Production of correct MWU lemmas is a prerequisite for the successful evaluation. Moreover, entries for morphological e-dictionary of MWUs can be produced only from correct MWU lemmas. Finally, as a byproduct of the whole process MWU inflectional classes for newly retrieved MWUs are obtained – they are derived directly from local grammars used for their extraction.

4.3 Evaluation of performance of MWU extraction

In order to evaluate our approach, we applied it to a collection of 74 papers in Serbian from the journal Infotheca.[6] The size of the corpus is

[6] Infotheca - Journal for Digital Humanities (http://infoteka.bg.ac.rs/index.php/en/infoteca)

272,557 simple word forms. Our procedure extracted from it 65,279 MWUs, 86.9% of them occurring only once, 7.9% occurring twice, 3.8% occurring 3 to 5 times and 1.9% with more than 5 occurrences.

The graph 3 (N2X) extracted 31% of all MWUs with frequency greater than 1. It is followed by graph 6 (NXN) with 26% MWUs, graph 4 (N4X) with 22%, graph 1 (AXN) with 16%, and the remaining six graphs with 6%. As to MWUs with frequency greater than 5, graph 1 (AXN) covers 31%, graph 3 (N2X) 25%, graph 6 (NXN) 22%, graph 4 (N4X) 17%, and the remaining six graphs 5%.

Extracted MWUs were manually evaluated on a subset of 690 entries. The evaluators checked 1) whether proposed lemmas were grammatically correct and 2) whether MWU terms belong to domain terminology, in this case library and information science, or to the general lexica.

For candidate ranking three measures were used: frequency, C-Value (Franzi et al., 2000) and log-likelihood (Dunning, 1993; Gelbukh et al., 2010).

For grammatical correctness best precision at rank *n* (P@n) measure is very high (figure 6) and independent of the ranking (the trend is flat).

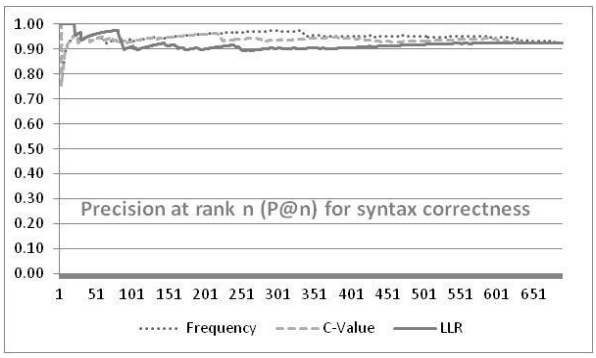

Figure 6: Precision at rank *n* for all evaluated term candidates for grammatical correctness.

In order to calculate the log-likelihood measure we used an excerpt from the general Corpus of Contemporary Serbian[7] that consists of 22 million simple word forms.

Figure 7 presents the precision at rank *n* for 690 evaluated term candidates for domain affiliation, where log-likelihood gave best results for precision at rank *n* (P@n) measured on a sorted list of candidates.

[7] The Corpus of Contemporary Serbian (http://www.korpus.matf.bg.ac.rs/)

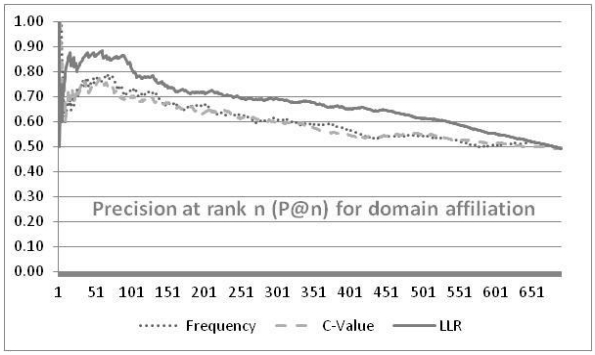

Figure 7: Precision at rank *n* for all evaluated term candidates for domain affiliation.

5 Discussion and Conclusion

The research outlined in this paper tackles the extraction of domain terminology and its integration into terminological dictionaries using lexical resources and local grammars. Results obtained by following this approach justify its basic assumption that the task of term extraction, both in the case of simple words and multi-word units, can be successfully accomplished combining existing e-dictionaries and FSTs. Moreover, lexical resources and local grammars alleviate the task of integrating the newly discovered terms into terminological dictionaries by simplifying the task of defining the proper inflectional class for new terms, a task extremely complex in case of morphologically rich languages such as Serbian. By implementing the procedure proposed within this paper we have considerably sped up the development of terminological dictionaries for Serbian.

Further research will address the integration of inflectional class prediction in existing software tools used for handling dictionaries developed at University of Belgrade and creation of a web tool that would support the entire procedure described in this paper. Production of dictionary entries in DELA format for verbs, akin to the one described for nouns, is also under consideration. A detailed evaluation will follow with the aim of further refinement of the presented procedure in order to reduce to the least possible extent the necessity for human intervention within the process of terminology acquisition and description. Our future work will be oriented towards usage of Web sites for evaluation of new term candidates (Robitaille et al., 2006).

Acknowledgement. This research was supported by the Serbian Ministry of Education and Science under the grant #47003 and Parseme COST action IC1207.

References

Ammar, C., Haddar, K., & Romary, L. (2015). Automatic Construction of a TMF Terminological Database Using a Transducer Cascade. *Proc. of Recent Advances in Natural Language Processing*. (pp. 17-23).

Baldwin, T., & Kim, S. N. (2010). Multiword expressions *Handbook of Natural Language Processing, second edition*. (267-292): CRC Press.

Cerbah, F., & Daille, B. (2007). A Service Oriented Architecture for Adaptable Terminology Acquisition. In Z. Kedad, N. Lammari, E. Métais, F. Meziane & Y. Rezgui (Eds.), *Natural Language Processing and Information Systems* (Vol. 4592: 420-426): Springer Berlin Heidelberg.

Courtois, B., Silberztein, M. (1990). Dictionnaires électroniques du français. Larousse, Paris.

Daille, B. (2000). Morphological rule induction for terminology acquistion. *Proc. of the 18th conference on Computational linguistics*- (Volume 1: pp. 215-221).

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

Enguehard, C., & Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics, 2*(1): 27-32.

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

Gelbukh, A., Sidorov, G., Lavin-Villa, E., & Chanona-Hernandez, L. (2010). Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In C. Hopfe, Y. Rezgui, E. Métais, A. Preece & H. Li (Eds.), *Natural Language Processing and Information Systems* (Vol. 6177, pp. 248-255): Springer Berlin Heidelberg.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1* (01): 9-27. doi:10.1017/S1351324900000048

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*: Faculty of Philology of the University of Belgrade.

Nakagawa, H., & Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology, 9*(2), 201-219.

Przepiórkowski, A., Degórski, Ł., & Wójtowicz, B. (2007). On the evaluation of Polish definition extraction grammars. *Proc. of the 3rd Language & Technology Conference*.

Quochi, V., Frontini, F., & Rubino, F. (2012). A MWE Acquisition and Lexicon Builder Web Service. *Proc. of COLING 2012* (pp. 2291-2306).

Ramisch, C., De Araujo, V., & Villavicencio, A. (2012). A broad evaluation of techniques for automatic acquisition of multiword expressions. *Proc. of ACL 2012 Student Research Workshop (1-6)*.

Robitaille, X., Sasaki, Y., Tonoike, M., Sato, S., & Utsuro, T. (2006). Compiling French-Japanese Terminologies from the Web. *Paper presented at the 11th Conference of the European Chapter of the Association for Computational Linguistics - EACL*.

Rodrıguez, F. M. B., Noya, E. D., Otero, P. G., Martınez, M. L., Mato, E. M. M., Rojo, G., Docıo, S. S. (2007). A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language. *Proc. of 3rd Language & Technology Conference*.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP *Computational Linguistics and Intelligent Text Processing* (1-15): Springer.

Savary, A. (2009). Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In S. Maneth (Ed.), *Implementation and Application of Automata* (Vol. 5642, pp. 237-240): Springer Berlin Heidelberg.

Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A. & Makowiecki, F (2012). SEJFEK—a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. *Proc. of Cognitive Aspects of the Lexicon (COGALEX-III)*. (pp. 195-214).

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. (pp. 36-44).