Automatic construction of a morphological dictionary of multi-word units

Cvetana Krstev, Ranka Stanković, Ivan Obradović, Duško Vitas, Miloš Utvić

Automatic construction of a morphological dictionary of multi-word units | Cvetana Krstev, Ranka Stanković, Ivan Obradović, Duško Vitas, Miloš Utvić | Lecture Notes in Computer Science 6233, Advances in Natural Language Processing, Proceedings of the 7thInternational Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 2010 | 2010 | |

10.1007/978-3-642-14770-8_26

http://dr.rgf.bg.ac.rs/s/repo/item/0000831

Automatic Construction of a Morphological Dictionary of Multi-Word Units

Cvetana Krstev[1], Ranka Stanković[2], Ivan Obradović[2], Duško Vitas[3], and Miloš Utvić[1]

[1] Faculty of Philology, University of Belgrade
[2] Faculty of Mining and Geology, University of Belgrade
[3] Faculty of Mathematics, University of Belgrade

Abstract. The development of a comprehensive morphological dictionary of multi-word units for Serbian is a very demanding task, due to the complexity of Serbian morphology. Manual production of such a dictionary proved to be extremely time-consuming. In this paper we present a procedure that automatically produces dictionary lemmas for a given list of multi-word units. To accomplish this task the procedure relies on data in e-dictionaries of Serbian simple words, which are already well developed. We also offer an evaluation of the proposed procedure on several different sets of data. Finally, we discuss some implementation issues and present how the same procedure is used for other languages.

Key words: electronic dictionary, Serbian, morphology, inflection, multi-word units, noun phrases, query expansion

1 Introduction

We have been developing morphological electronic dictionaries of Serbian for natural language processing for many years now. Our e-dictionaries follow the methodology and format known as DELAS/DELAF, which is presented for French in [1][4]. Serbian e-dictionaries of simple forms have reached a considerable size: they have a total of 122,000 entries [2]. Although we are continually enlarging our e-dictionaries of simple words, we have taken a further step towards tackling the problem of multi-word units (MWUs). In an introduction in [3], supported by comprehensive references, Savary states that MWUs are hard to define controversial linguistic objects. Nevertheless, it seems that many authors agree that MWUs: (a) are composed of two or more graphical words; (b) show some degree of morphological, syntactic, distributional, or semantic non-compositionality; and (c) have unique and constant references. A deeper discussion of the nature of MWUs is beyond the scope of our paper. When dealing with MWUs we are taking a pragmatic approach and consider MWUs to be sequences of graphical units that for some reason have to be described and processed as a unit.

[4] A comprehensive bibliography on e-dictionaries for other languages developed using the same methodology is given at `http://www-igm.univ-mlv.fr/~unitex/`.

For some productive classes of MWUs, like different types of numerals and named entities (time and duration, measures and currencies), we have developed finite-state transducers (FSTs) that rely on morphological e-dictionaries of simple words to model these MWUs correctly [4]. When applied to a text in automatic text analysis these FSTs associate recognized MWUs with lemmas as well as with appropriate grammatical categories. Both the associated lemmas and grammatical categories are in the same format as the one provided by dictionaries of simple words.

Some other MWUs, that are idiosyncratic in nature, ask for a different description. Namely, these MWUs cannot be described by FSTs - they have to be listed in an e-dictionary in a similar way and for similar reasons as simple words. That means that some regular form, or a lemma, has to be listed in a DELAC dictionary, together with some additional information that would enable the generation of all inflected forms, that is, entries for a DELACF dictionary. Several questions arise here: (a) what should this regular form listed in a dictionary of lemmas be; (b) what additional information is necessary; and (c) how the generation of all forms is to be performed. When trying to find the answers to these questions one has to keep in mind that the graphical units composing a MWU are themselves simple words that have their own inflectional behavior. However, simple words that represent components of a MWU do not inflect freely - they have to conform to some combining rules. In the case of Serbian, these rules can be rather complex, since simple words constituting a MWU, like adjectives and nouns, can inflect in several grammatical categories.

For instance, *civilni vojni rok* 'civil military service' is a multi-word noun that inherits its gender from the constituent noun *rok* (masculine in this case), and it inflects for case, but it does not inflect for number (although the simple word *rok* does). The adjectives *civilni* and *vojni* agree with the noun *rok* in number, case, gender and animacy, but the comparative and indefinite forms of these adjectives are not used in this MWU. It is clear from this example that the combining rules for Serbian MWUs are by no means trivial. After considering several options, we concluded that Multiflex, a finite-state tool for MWUs, developed by A. Savary [5], suits our needs best. This tool supports finite-state transducers that can model a number of combining conditions. As for the inflection of constituent simple words, it relies completely on FSTs for the inflection of simple words. This way, the generation of all inflected forms with Multiflex is performed with two types of FSTs: for inflection of simple words and for inflection of MWUs. To that end, a lemma for each simple word constituent that inflects in a MWU has to be provided, as well as values of all grammatical categories of forms appearing in the MWU lemma (its regular or dictionary form). For the given example, the entry in DELAC dictionary is:

civilni(civilni.A2:adms1g) vojni(vojni.A2:adms1g) rok(rok.N81u:ms1q),NC_AXAXN1

The information given in this entry allows automatic production of all 26 inflected forms for the DELACF dictionary as, for example, the entry for the singular dative case:

civilnom vojnom roku,civilni vojni rok.N:ms3q

Although the Multiflex approach is theoretically well-founded, easy to understand and apply, and it successfully solves many problems of MWU inflection for Serbian, it is obvious from the given example that the production of a single DELAC entry is a very tedious task. As a matter of fact, we initially produced only 30 DELAC entries from scratch. Realizing that additional information within DELAC entries (everything between the parenthesis) in most cases already exists in dictionaries of simple words (DELA) we decided to develop a module for our lexical resources management tool LeXimir, an enhancement of its predecessor WS4LR [6], that would help in obtaining this information. However, due to homography of forms and homonymy of simple word entries, the developer of a DELAC entry still needs to choose between several options offered by a dictionary look-up provided by this tool. For the above example, the choice had to be made three times for the information about forms (like 'adms1g' for the first component) and once for the simple word entry (because there are two entries for *rok* in the Serbian DELAS: one for 'service' and one for 'rock music'). Following this approach, only 3195 DELAC entries were produced in the past three years, which we found very ineffective.

In a comprehensive analysis of several tools for MWU inflection description [3] the author mentions only one system, *FASTR*, that supports automated MWU lexicon creation [7]. Since this system is based on an approach very different from DELA methodology, we developed our own procedure for automatic construction of DELAC entries from a given list of MWUs. For an item such as *civilni vojni rok* this procedure produces the aforementioned MWU lemma. However, the produced list of lemmas has to be manually checked and some decisions still have to be made even when the procedure offers only one candidate MWU lemma. Thus we have to stress that our dictionaries remain handcrafted resources (as per categorization in [8]) in terms of the information they offer and the methodology used to produce them (without statistical engineering).

2 Analysis of Initial Data

We based the development of the automated procedure for DELAC construction on the initial dictionary of MWUs, which contained 3195 lemmas covering different part of speech as presented in Table 1: nouns, adjectives, adverbs, conjunctions, prepositions, interjections. As only MWU nouns and adjectives inflect, they have an inflectional class code assigned to them in DELAC. Each inflectional class is associated with one inflectional transducer (as described in [9]) that controls the production of all inflected forms. The forms/lemmas ratio in Table 1 shows that a direct production of DELACF dictionary is out of the question for Serbian, though it may seem possible for some other languages. For instance, for English this ratio for MWU nouns is 1.90, whereas for French it is 1.29. The calculation is performed on the basis of dictionaries described in [10] and [11] that are part of the standard distribution of Unitex [12], a corpus processing system based on the finite-state technology.

Table 1. Initial content of the Serbian morphological dictionary of MWUs

POS	lemmas	forms	forms/lemmas	Inflectional classes	Inflectional super-classes
Nouns	2571	49792	19.4	67	28
Adjectives	207	21045	101.7	16	9
Other	415				

Some inflectional transducers are frequently needed (like NC_AXN, which is used for MWUs consisting of an adjective followed by a noun, in our case 1208 MWUs), while some apply to a single MWU. In order to design our automated procedure we grouped all inflectional transducers into equivalence classes or super-classes: a super-class consists of all transducers that use the same form of a MWU lemma, that is, the same information for the production of inflectional forms. This is also reflected in the convention we used for naming the inflectional transducers: A stands for an adjective constituent, N stands for a noun constituent, X stands for a constituent that does not inflect (including a separator), while some additional digits and letters may be added to differentiate transducers. This is illustrated in Table 2 by six inflectional transducers all belonging to one super-class NXN and used for the inflection of MWUs consisting of a noun followed by another noun, where both nouns inflect and must agree in basic grammatical categories.

It should be noted that MWUs sharing the same inflectional class (or super-class) do not necessarily have the same syntactic structure and vice versa. For instance, *film u boji* 'color movie' and *bolest ludih krava* 'mad cows disease' share the same class NC_N4X, although the first MWU consists of a noun followed by a prepositional phrase and the second of a noun followed by a phrase in the genitive. However, from the inflectional point of view they both behave in the same way: only the first component inflects, whereas the second and the third do not. On the other hand, *predsednik države* 'president of the state' and *advokat odbrane* 'attorney of the defense' both consist of a noun followed by a component in the genitive case, but their inflectional behavior is different: in the first MWU the second component can change in number whereas in the second MWU it does not inflect. For that reason they belong to two different classes (and super-classes), NC_NXN4 and NC_N4X respectively.

In order to formulate our strategy for the production of MWU lemmas we analyzed the data in the initial dictionary looking for useful information. We identified the additional information assigned to components of MWUs belonging to a particular inflectional class, and vice versa, we identified inflectional classes associated with the same additional information. For instance, the class NC_NXN2m explained in Table 2 is associated with only one combination of grammatical categories, characterized by the fact that the components differ in gender. On the other hand, some combinations of grammatical categories like [ms1q,ms1q] (two masculine inanimate nouns in the nominative singular) were associated in DELAC with three classes: NXN, NXN2, and NXN3 (see Table 2), the first one being the most frequent.

Table 2. The inflectional classes for MWUs belonging to the super-class NXN

Infl. class	Example	Translation	Explanation
NC_NXN	lekar akušer	obstetrician	Both components inflect and agree in case and number
NC_NXNF	kit ubica	killer whale	The gender of the second component changes in plural
NC_NXN3	kamen-temeljac	foundation stone	The separating hyphen can be omitted
NC_NXN2	Kongo Brazavil	Congo-Brazzaville	Neither of the components inflects for number
NC_NXN4	predsednik države	president of the state	The first component inflects for case and number; the second may or may not inflect for number
NC_NXN2m	Kneževina Monako	Principality of Monaco	The second component does not necessarily inflect

On the basis of this information we got the general idea which combinations of atomic data can we expect to be extracted from dictionaries of simple words and how they combine with inflectional classes. However, the obtained information was incomplete. For components that do not inflect for a certain MWU there was no additional information in the MWU lemma. For such components this information was subsequently searched in dictionaries of simple words. It was thus possible to establish that in *naliv-pero* 'ink pen' the first component was not in the dictionary of simple words, in *bruto plata* 'gross income' the first component was an adjective that does not inflect, in *mini-suknja* 'mini-skirt' the first component was a prefix. All three examples belong to the same super-class 2XN. There are still cases when no additional information, whether extracted from MWU lemmas themselves or subsequently from electronic dictionaries, can help in deciding in favor of one inflectional class over other possibilities. For instance, this is the case for *krem-karamel* 'caramel cream' and *kamen-temeljac* 'foundation stone': both components in these MWUs are masculine inanimate nouns in the nominative singular, but in the first one *karamel* is the head and the first component does not inflect (super-class 2XN), while in the second example *kamen* is the head and both components inflect (super-class NXN).

3 Description of the Strategy

The purpose of the analysis performed in the first step was not to produce a strategy for construction of MWU lemmas automatically — it was rather used as a reference during the manual production of a strategy in the form of XML documents. The schema of these documents is presented in Figure 1. Our strategy consists of two XML documents, one for MWU nouns, and the other for MWU adjectives. Each XML document consists of a sequence of rules that are grouped by the number of components in MWUs. Each rule states the conditions that a MWU and its components have to satisfy in order to be placed in a particular inflectional class and to have a particular MWU lemma assigned to them. In order to make our strategy easier to produce and maintain, conditions for some

rules were grouped into general and specific conditions, where specific conditions are simply additions to general conditions. One rule will illustrate this:

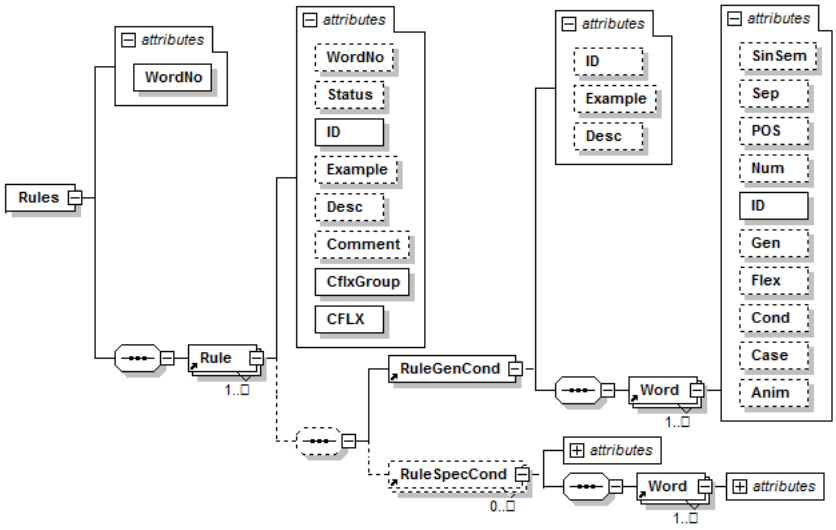

Fig. 1. The XML Schema for a strategy document

```
<Rule ID="2" CFLX="NC_AXN3" CflxGroup="AXN">
 <RuleGenCond>
  <Word ID="1" POS="A" Flex="true" Case="1" Anim="$a" Gen="$g"/>
  <Word ID="2" POS="N" Flex="true" Case="1" Anim="=$a" Gen="=$g"/>
 </RuleGenCond>
 <RuleSpecCond ID="1" Example="Ajfelova kula">
  <Word ID="1" Num="s" Cond="$PRE"/>
  <Word ID="2" Num="s"/>   </RuleSpecCond>
 <RuleSpecCond ID="2" Example="poljski radovi">
  <Word ID="1" Case="1" Num="p"/> <Word ID="2" Case="1" Num="p"/>
 </RuleSpecCond>
 <RuleSpecCond ID="3" Example="elektronsko poslovanje">
  <Word ID="1" Case="1" Num="s"/>
  <Word ID="2" Case="1" Num="s" SinSem="+VN,+Coll,+HumColl"/>
 </RuleSpecCond>
</Rule>
```

This rule classifies a MWU in the inflectional class NC_AXN3 associated with adjective-noun MWUs that do not inflect for number (super-class AXN) if the following conditions are satisfied:

a) General conditions: the first component is an adjective, the second is a noun, both components are listed in the nominative case, they agree in gender and animacy (whichever they are);
b) Additional specific conditions:

- both components are in singular and the first starts with upper-case (*Ajfelova kula* 'Eiffel tower'), or
- both components are in plural (*mirovne snage* 'peace forces'), or
- both components are in singular and the second component is a verbal noun or a collective noun (*belo roblje* 'white slaves').

The number of rules for nouns and adjectives are listed in Table 3. Some rules have no special conditions while some have up to ten special conditions.

Table 3. Overview of rules for nouns and adjectives with 2 to 7 components

Components	**Rules for nouns**		**Rules for adjectives**	
	(general)	**(special)**	**(general)**	**(special)**
2	26	59	5	9
3	25	85	8	8
4	14	50	5	5
5	6	54	2	2
6	4	29		
7	1	9		
Total	76	286	20	24

The rules are applied in the sequence in which they appear in the XML document, which means that if more than one rule can apply for a particular MWU, then more MWU candidate lemmas will be offered in the order of the application of rules. Hence, the order of rules in the strategy is of importance. For instance, there are three rules that can apply to the noun-noun inflectional class NC_NXN:

1. The first rule is 9^{th} in the sequence of rules and its conditions are very strict: both components have to agree in all four grammatical categories: number, case, gender and animacy (e.g. *lekar akušer* 'physician obstetrician');
2. The second rule is 21^{st} in the sequence and its conditions are less strict: the components do not have to agree in animacy (e.g. *biljka*(q) *mesožderka*(v) 'plant carnivore');
3. The third rule is 22^{nd} in the sequence and allows the components to disagree in gender (e.g. *roman*(m) *reka*(f) 'novel fleuve').

As a consequence, for noun-noun MWUs in which both nouns do not agree in all grammatical categories, the other rule producing the NC_2XN class will have precedence, as this class is, in general, more frequent.

There are some inflectional classes for which no rule can be applied: we identified 14 such classes for nouns and 2 for adjectives. This is usually related to optional use or omission of a hyphen, and in these cases the strategy offers a very similar class. There are also some classes that are related to only one very specific MWU, so we have not even tried to formulate a rule for them, e.g. the class NC_2XN3XN1 for *Sao Tome i Principe*. However, all super-classes are covered by rules.

4 Analysis and Evaluation of the Strategy

We have performed the analysis and evaluation of our strategy in two steps. In the first step we applied the strategy to the same data that we used to produce it, that is our initial DELAC dictionary. In the second step, we applied it to several lists of MWUs that we have collected from various sources.

After applying our strategy to the initial DELAC dictionary containing 2571 nouns and 207 adjectives the obtained results were manually validated. When assessing the success of the strategy we adhered to the following:

a) If for a given MWU one of the rules produced the correct lemma and assigned the correct inflectional class, we considered the strategy as successful;
b) If for a given MWU none of the rules satisfied a), but one of the rules produced the correct lemma, although the assigned inflectional class was not correct, whereas the assigned super-class was correct, we considered the strategy as partially successful;
c) If for a given MWU none of the rules satisfied a) or b) we considered that the strategy failed for that MWU;
d) For each lemma where either a) or b) applied, we also determined the rank of the accepted solution: the higher on the list of offered solutions, the more favorable the accepted solution.

Table 4 gives percentages of successfully produced noun and adjectives entries (case a), partially successful results (case b) and failures (case c), and also indicates the rank of these results (rank 0 means no solution offered). A failure means that either no solution was offered or none of the solutions was classified as a) or b). In either case, the reason was that the strategy failed to cover a particular MWU structure (in 52 cases, or 20% of all failures) or a MWU component that inflects was not in the dictionary of simple words (in 255 cases, or 80% of all failures). The latter case occurred frequently due to MWUs representing proper names, where components are often not words in Serbian, e.g. *Dar es Salam*, or due to the fact that some words are used only in MWUs (like *domali* in *domali prst* 'next to little (ring) finger') and are thus not listed in dictionaries of simple words. The lowest rank of a successful result was 10 for nouns, and 3 for adjectives.

Table 4. Evaluation of the strategy applied to the initial DELAC dictionary

Rank	**Nouns**				**Adjectives**			
	a)	**b)**	**c)**	**Total**	**a)**	**b)**	**c)**	**Total**
0			5.50%	5.50%			2.93%	2.93%
1	63.74%	12.69%	6.28%	82.71%	37.56%	15.61%		53.17%
2	6.56%	0.62%	0.08%	7.26%	36.59%	4.39%		40.98%
3	2.22%	1.21%		3.43%	2.93%			2.93%
4-10	0.90%	0.20%		1.10%				0.00%
Total	**73.42%**	**14.72%**	**11.87%**	**100.00%**	**77.07%**	**20.00%**	**2.93%**	**100.00%**

In the second step we applied our strategy for nouns to several different lists of MWUs. We have not applied our strategy for adjectives in this step

simply because we have not collected enough new adjectives. Our list of new MWU nouns came from several different sources: the official list of MWU names of settlements in Serbia (236), MWUs extracted from a log file of a Serbian professional journal that deals with economic issues (162), from Verne's novel *Around the world in 80 days* (114), from the explanatory dictionary of Serbian, under the letter R (604). As the analysis in the first step indicated that MWU proper names produce in general worse results, we decided to separate these lists in two groups. After removing those already in DELAC we got a list of MWU toponyms (206) and a list of MWU common nouns (784).

Table 5 shows that in the case of common nouns, for only 3.62% items on the list (28 items) no satisfactory solution (a or b) was offered. For toponyms this percent is much higher (38.61%), accounting for 78 items on the list. In the case of toponyms all cases of failure are due to the fact that components of toponyms were not simple words in Serbian (e.g. *Feja* in *Kriva Feja*). The lowest rank of a successful result was 6 for common nouns, and 2 for toponyms.

Table 5. Evaluation of the strategy for nouns applied on a list of MWU common nouns and compound toponyms

Rank	**Common nouns**				**Toponyms**			
	a)	**b)**	**c)**	**Total**	**a)**	**b)**	**c)**	**Total**
0			1.68%	1.68%			24.75%	24.75%
1	80.23%	10.08%	1.94%	92.12%	48.02%	3.47%	13.86%	65.35%
2	4.39%	0.26%		4.78%	8.91%			8.91%
3	1.29%	0.13%		1.42%				0.00%
4-6				0.00%	1.00%			1.00%
Total	**85.92%**	**10.47%**	**3.62%**	**100.00%**	**57.92%**	**3.47%**	**38.61%**	**100.00%**

Row 2 in Table 6 shows that less rules were used for common nouns in the second step than for nouns in the initial DELAC in the first step. This is probably due to the fact that while building our initial DELAC and the inflectional classes we tried to find various structurally different examples. Row 3 shows that for all subsets (except adjectives) the number of rules that were not used is greater than the number of used rules, namely, many rules were added to the strategy upon analogy with other rules. As the number of rules does not affect the effectiveness of the procedure (except the processing time to a small degree) we believe that unused rules should not be removed because they may prove useful in the future. Indeed, the subset "common nouns" used seven rules that were not used for the initial DELAC nouns, while "toponyms" used one new rule.

The rules used most frequently for nouns (row 5) are rules pertaining to adjective-noun MWUs, which are also the most frequent in the dictionary. The rules that failed each time they were used (row 7) are obvious candidates for deletion from the strategy. Rules that were more unsuccessful than successful (row 8) should probably also be reconsidered, as for instance, by reinforcing the conditions, if possible.

The average number of solutions per item is rather low (row 9), ranging from 1.4 for the "common nouns" to 3.7 for "toponyms". In some cases much

larger sets of solutions were offered. The leader is *Velika Plana* (a small town in Serbia) with 52 solutions offered. Such a high number of solutions occurred due to the homography of both components. However, even in this case the correct solution had a high rank: namely, the second solution for *Velika Plana* was correct. It may seem reasonable to exclude some dictionaries of simple words from this procedure, for instance, dictionaries of personal names, in order to alleviate similar problems. However, that might not be such a good idea: these very dictionaries successfully processed many items, among them three very specific names of small towns in Serbia named after famous Serbian poets and politicians: *Aleksa Šantić*, *Svetozar Miletić*, *Jaša Tomić*.

Table 6. Strategy rules performance data for subsets: **N** (initial DELAC nouns), **A** (initial DELAC adjectives), **CN** (additional common nouns), **T** (additional toponyms)

	N	**A**	**CN**	**T**
1. Items	2571	207	784	206
2. Rules applied	85	19	56	33
3. Rules not applied	201	5	230	253
4. Applications of rules	4083	434	1060	769
5. Most frequently used rule	NC_AXN	AC_2X2	NC_AXN	NC_AXN3
(number of times applied)	1499	108	589	144
6. Absolutely successful rules	26	9	19	1
7. Absolutely unsuccessful rules	9	1	16	26
8. Rules more unsuccessful than successful	36	6	23	29
9. Solutions/item	1.6	2.1	1.4	3.7
10. Maximum solutions per item	38	6	19	52
11. Success 1^{st} solution	80.64%	54.77%	91.85%	68.43%
12. Success 2^{nd} solution	7.58%	42.21%	4.73%	11.84%
13. Success 3^{rd} solution	3.62%	3.02%	1.44%	0.0%

Successful outcomes, that is solutions classified as a) or b) offered at the first, second or third place (rows 11, 12 and 13) were counted for cases where the strategy offered at least one solution (not necessarily an acceptable one). These data show that for nouns, in general, if a solution is offered, it is very likely that the first one will be correct. For adjectives, the most likely solution will be among the first two offered.

5 Implementation and Usefulness

Our procedure for automatic production of DELAC entries is a module of the LeXimir tool [6], which is written in C# and operates on the .NET platform. It supports development, maintenance and exploitation of various resources: e-dictionaries, wordnets, and aligned texts. A user of this tool need not use all of these resources, or even possess them, but those that exist are visible in all modules and can be exploited in a useful way.

The e-dictionaries of simple and MWU words that we develop using LeXimir are used primarily within the Unitex system. As Unitex is open source software

distributed under the terms of LGPL, we easily incorporated its modules in LeXimir for many tasks that involve manipulation of e-dictionaries, including dictionary look-up used in the module for (automated) production of DELAC entries. To manipulate the strategy in the form of XML documents our tool relies on W3C standard languages XQuery and XSLT supported by .NET.

The user interface of the module for automatic production of DELAC lemmas is very friendly. A user can choose files with lists of MWUs and a strategy, and results are presented in a form of a table in which the user has only to check the correct solutions upon which a list of DELAC entries is produced. Various debugging tools and preference selections are at the user's disposal.

It has already been shown that LeXimir can be used for languages other than Serbian and English. Our new module for production of DELAC entries can also be successfully applied without any modification to other languages that have dictionaries of simple words in DELA (thus, directly supported by Unitex); naturally, corresponding XML documents representing the strategy for a particular language need to be created. However, the system can be easily modified to support other formats of simple words dictionaries because only the dictionary look-up module has to be changed. The experiment is already in progress for Polish, for which Multiflex is used for inflection of MWUs but simple word dictionaries are handled in a different, database environment [13].

Another tool WS4QE (shortened for Work Station for Query Expansion) was developed on basis of LeXimir, and it enables expansion of queries submitted to the Google search engine [6]. Integrated lexical resources enable modifications of user queries for both monolingual and multi-lingual search. The main feature of WS4QE is that it enables inflection of simple words and MWUs supplied as key-words to Google. Again, Unitex and Multiflex are used for inflection. However, WS4QE goes one step further: for free phrases supplied as key-words having a structure covered by a MWU inflectional class, the tool uses our strategy and acts upon the first offered solution, which is the correct one in most cases.

6 Conclusions

The results that we have presented justify the efforts invested in designing our procedure because it allows for massive production of DELAC entries. We have already prepared a list of 25,000 MWUs extracted from the Serbian explanatory dictionary that we hope to be able to process in a few months. One important thing that remains to be done is the addition of semantic and/or domain markers to MWU lemmas, which have so far been systematically added only for proper names following the approach suggested in [14]. We are considering several solutions, including one proposed in [15].

We envisage further development of our procedure. We would like to allow MWUs to be components of other MWUs (and components of free phrases as well). This would keep the number of possible structures low, and consequently reduce the number of inflectional classes and the number of rules in the strategy.

For instance, a lemma for a MWU from the beginning of this article *civilni vojni rok*, could in this case be:

civilni(civilni.A2:adms1g) {vojni rok}(vojni rok.NC_AXN:ms1q),NC_AXN

That is, its second component could be a MWU itself (*vojni rok* 'military service') and not a simple word and it would remain in the most frequent adjective-noun class. This approach is already implemented in Multiflex but not in Unitex. However, DELAC entries that are already produced need not be revised.

References

1. Courtois, B., Silberztein, M.: Dictionnaires électroniques du français. Larousse, Paris (1990)
2. Krstev, C.: Processing of Serbian - Automata, Texts and Electronic Dictionaries. Faculty of Philology, University of Belgrade, Belgrade (2008)
3. Savary, A.: Computational Inflection of Multi-Word Units - A Contrastive Study of Lexical Approaches. Linguistic Issues in Language Technologies **1** (2008)
4. Krstev, C., Vitas, D.: Finite State Transducers for Recognition and Generation of Compound Words. In Erjavec, T., Žganec Gros, J., eds.: IS-LTC 2006, Ljubljana, Slovenia, Institut "Jožef Stefan"(2006) 192–197
5. Savary, A.: Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In: CIAA. (2009) 237–240
6. Krstev, C., Stanković, R., Vitas, D., Obradović, I.: The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines. In: 6th LREC, Marrakech, Marocco (2008)
7. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. MIT Press (2001)
8. Laporte, E.: Lexicons and Grammars for Language Processing: Industrial or Handcrafted Products? In Rezende, L.M., da Silva, B.C.D., Barbosa, J.B., eds.: Léxico e gramática: dos sentidos à construção da significação. Trilhas Lingüísticas, 16. São Paulo: Cultura Acadêmica (2009) 51–84
9. Krstev, C., Vitas, D., Savary, A.: Prerequisites for a Comprehensive Dictionary of Serbian Compounds. In Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T., eds.: FinTAL. Volume 4139 of LNCS., Springer (2006) 552–563
10. Savary, A.: Recensement et description des mots composés - méthodes et applications. PhD thesis, Université de Marne-la-Vallée (2000)
11. Courtois, B., Garrigues, M., Gross, G., Gross, M., Jung, R., Mathieu-Colas, M., Silberztein, M., Vivès, R.: Dictionnaire électronique des noms composés DELAC : les composants NA et NN. Technical Report 55, LADL, Université Paris 7 (1997)
12. Paumier, S.: Unitex 2.1 User Manual, http://www-igm.univ-mlv.fr/unitex/UnitexManual2.1.pdf. (2008)
13. Wolinski, M., Savary, A., Sikora, P., Marciniak, M.: Usability Improvements in the Lexicographic Framework Toposlaw. In Vetulani, Z., ed.: 4th LTC, Poznań, Poland, IMPRESJA Widawnictwa Elektroniczne S.A. (2009)
14. Grass, T., Maurel, D., Piton, O.: Description of a Multilingual Database of Proper Names. In Ranchod, E., Mamede, N.J., eds.: PorTAL. Volume 2389 of LNCS., Springer (2002) 137–140
15. Elia, A.: The Electronic Thematic Linguistic Atlases (Atlanti Linguistici Tematici Informatici - ALTI), Atlas DICoMP - Dizionario delle parole composite. (http://www.ricercaitaliana.it/prin/unita_op_en-2005109535_003.htm)