Prerequisites for a Comprehensive Dictionary of Serbian Compounds

Cvetana Krstev[1], Duško Vitas[2], and Agata Savary[3]

[1] Faculty of Philology, University of Belgrade, Belgrade
[2] Faculty of Mathematics, University of Belgrade, Belgrade
[3] Computer Science Laboratory, François-Rabelais University of Tours, Blois Campus

Abstract. The paper describes the steps that were undertaken in order to start the production of a comprehensive morphological dictionary of compounds for Serbian. First, the classes of multi-word expressions were determined that were to be covered by the dictionaries. In the next step the useful sources of compounds were detected. The retrieved compounds were then classified according to their inflectional properties. The recently developed special finite state transducers were constructed for each of these classes which produce all the variants and morphological forms for the compounds of the class. Finally, the software module was developed that facilitates the production of the dictionary of compound lemmas with all the necessary information in the required format.

1 Introduction

The morphological dictionary of the simple words of Serbian is being developed following the LADL methodology ([2]) during the last decade ([21]). The dictionaries have reached such a size that enables the effective processing of Serbian texts: the dictionaries of general lexica having 80,000 lemmas (yielding 1,100,000 word forms) are supplemented by special dictionaries of proper names that have 29,000 lemmas (yielding 185,000 word forms). The comprehensiveness of these dictionaries enables the text coverage that leaves from 1 to 5% of unrecognized words.

The next step in the development of the lexical resources for Serbian is to produce the dictionaries of compounds in the same format (called DELAC, cf [18]) . For Serbian, as for the other Slavic languages, this task is not easy to accomplish. The characteristics of Serbian that make it particularly demanding are:

1. *Phonologically based orthography*, the consequence of which is that a considerable number of morphophonemic processes are reproduced in written texts.
2. *Transcription* of all foreign proper names according to the Serbian orthography.
3. *The rich morphological system*, which is reflected both on the inflective and derivational level ([20]).

4. *Free word order* of sentence constituents, *special placement of enclitics*, and *complex agreement system* ([1]).

As a consequence, it is not recorded in the scientific literature, to our best knowledge, that a comprehensive morphological dictionary of compounds has been developed for some Slavic language.

The question often arises how many compounds exist in a language. The French DELACF dictionary of compound word forms from 2002 contains 248,885 entries compared to the 746,214 entries in the DELAF dictionary of simple word forms. For the languages that are starting to build such a dictionary one answer can be found in the Wordnets for particular languages. For instance, there are 12636 compound literals out of 44910 in Bulgarian Wordnet (28.14%) and respectively 4074 such literals out of 18390 existing in Serbian Wordnet (SWN) (22.15%) ([8]). It is to be supposed that in a more developed Wordnet, in which more synsets belonging deeper in the hypernym/hyponym hierarchy would be added, the contribution of the compounds would be even greater. For instance, the synset <trophy:2, prize:3> is in the eighth level node in a hypernym/hyponym branch of the Princeton Wordnet 2.0 (PWN), and its three hyponyms <bronze medal:1>, <silver medal:1>, and <gold medal:1> are all represented by compounds. The same situation exists for the corresponding synsets in the SWN.

2 Definition of Compounds

The notion of a compound is controversial among both linguists and NLP-researchers ([3], [4], [5]). In [18] compounds are defined as sequences of simple words (which are strings of alphabetic characters of a given language) that show some degree of non-compositionality from the morphological, distributional, syntactic or semantic point of view.

The limit between noun compounds and free nominal groups is not always easy to establish. For instance the noun phrase *plavo nebo* 'blue sky' is a frequent one (35 occurrences in the Corpus of Contemporary Serbian (CCS) [21]) since one often describes sky as blue; however, one can not treat it as a compound since it does not represent a new concept. The noun phrase *plava grobnica* 'blue burial chamber', however, does not represent the burial chamber that is blue but is used to refer to the burial place of those that died on the sea. The noun phrase *plavi šlemovi* 'blue helmets' referring to the UN peace forces illustrates some other compound features: *šlem* 'helmet' represents an artifact, while *plavi šlemovi* represents an organization. This example also shows that new compounds emerge in a language regularly, and it cannot be known in advance how long they will last.

The structure of a compound is stricter then that of a free noun phrase: compounds usually do not allow a change of the word order or insertions ([4] talks about the degree of "fixedness" which is the higher the more syntactic transformations are forbidden for the given phrase). In Serbian, the free noun phrase *plavi šlemovi* could be expressed equivalently as *šlemovi plavi*; the latter phrase,

however, cannot be used to denote the UN peace forces. Also, the presence of the inserted adjective in *plavi zaštitni šlemovi* 'blue safety helmets' indicates that the literal meaning is used. Although this is in general true, it does not mean that there are no exceptions: for instance, *žuta tampa* 'yellow journalism' is a compound, and consequently, the occurrence *verska žuta štampa* would refer to the religious journalism of the sensationalist kind. However, the CCS records also *žuta verska štampa* which shows that in this case the adjectives can be freely distributed as in a free noun phrase.

Compounds should also be distinguished from verb phrases. For instance, *plavi dres* 'blue gym suit' is sometimes used by sport journalists to refer to the Serbian national team, regardless of the sport in question. However, it cannot be regarded as the synonym of *reprezentacija* 'national team' since these two are not interchangeable. Namely, one cannot rephrase *Jugoslovenska košarkaška reprezentacija nije otputovala na Olimpijske igre* 'Yugoslav national basketball team did not leave for the Olympic Games' by **Jugoslovenski košarkaški plavi dresovi nisu otputovali na Olimpijske igre*. The minute analysis of the usage of the expression *plavi dres* shows that it is used only in a restricted number of phrases, such as *igrati za plavi dres* 'to play for the blue gym suit', where the verb *igrati* 'to play' can be replaced only by a few other (*odigrati*, *zaigrati* 'perfective forms of to play', *voziti* 'to drive', *nositi* 'to wear', *zaslužiti* 'to deserve', etc.). The expression *plavi dres* can be replaced by *reprezentacija*, but only if the sentence is rephrased: *Poslednju utakmicu Žućko je odigrao u plavom dresu koji je nosio celu deceniju* 'Žućko has played his last game in a blue gym suit that he wore for a whole decade' can be changed to *Poslednju utakmicu Žućko je odigrao za reprezentaciju za koju je igrao celu deceniju*. This cannot be treated as a compound and it will be treated as phrase.

One of the roles compounds have in text processing is in disambiguation since in many cases compounds can be unambiguously recognized. That is, they invalidate the interpretations obtained by tagging the word forms that are their constituent parts. In Serbian, the most convincing is the case of *Crne Gore*, the genitive case form of *Crna Gora* 'Montenegro'. When dictionaries of simple word forms are applied to this sequence the following result is obtained:

```
({crne,crn.A+Col:aemp4g:aefs2g:aefw2g:aefw4g:aefp1g:aefp4g:aefp5g} +
 {crne,crneti.V547+Imperf+It+Iref+Ref+Ek:Pzp:Ays:Azs} +
 {crne,crnjeti.V747+Imperf+It+Iref+Ref+Ijk:Pzp})
({gore,gora.N:fs2q:fw2q:fw4q:fp1q:fp4q:fp5q} + {gore,gore.ADV} +
 {gore,goreti.V544+Imperf+It+Iref+Ek:Pzp:Ays:Azs} +
 {gore,gorjeti.V744+Imperf+It+Iref+Ijk:Pzp} +
 {gore,rdjav.A:bemp4g:befs2g:befw2g:befw4g:befp1g:befp4g:befp5g:bens1g...} +
 {gore,zao.A:bemp4g:befs2g:befw2g:befw4g:befp1g:befp4g:befp5g:bens1g...})
```

The word form *crne* obtains 11 grammatical interpretations for 3 different lemmas, while *gore* obtains 31 grammatical interpretations for 6 different lemmas. These are all the cases of a "false ambiguity" ([12]) since a human reader does not see them as such; if written in this way, with both simple word forms with initial

capitals, it represents the Republic Montenegro, and it can be unambiguously tagged: `Crne Gore,Crna Gora.AN+C+Nprop+Top+Dr:fs2q`

3 Collecting

The compounds that will be covered by our Serbian DELAC can be grouped in various Parts-of-Speech. In Serbian the compounds that do not inflect are compound prepositions (*bez obzira na* 'regardless of'), conjunctions (*kao da* 'as if'), interjections (*blago tebi* 'lucky you'), and adverbs (*od srca* literally 'from heart' meaning 'willingly', *iz dana u dan* 'day in day out'). The compound numerals occur often in texts (*dvadeset i pet miliona* 'twenty five millions'), but as they are built in regular way from a small number of constituents, they are usually not part of a dictionary but are recognized using other tools, such as FSTs. The same is valid for many adverbial phrases, as *januara prošle godine* 'in January last year' and they are treated in the similar way. The compounds that inflect can be categorized as adjectives (*kulturno-umetnički* 'cultural and artistic') and nouns (*general pukovnik* 'general colonel', *ministar spoljnih poslova* 'minister of the foreign affairs').

There exist many approaches dedicated to manual, semi-automatic or automatic extraction of compounds of various types such as frozen expressions, complex terms (see [6] for a comparative study of some of them), multi-word named entities (e.g. [13]), etc. We know of no such method for Serbian. Some extraction systems, based mainly on statistical estimation of token co-occurrences, are meant to be language-independent. One such system has been used for term extraction from Serbian texts in restricted domains but the results were not very promising ([14]).

As stated the section 1 Wordnet can be regarded as a valuable source of potential compounds. However, not all literals in Wordnet that contain non-alphabetic characters are compounds, since quite a number of them are just descriptions of some concepts. For instance, in PWN the synset <group action:1> is defined as an 'action taken by a group of people'. The corresponding synset in SWN is <grupna akcija:X> and although the English literal may be regarded as a compound, the Serbian one can hardly be.

Another source of compounds is the list of unknown words produced during the lexical analysis since the constituents of various compounds can be found in it. For example, in *akten-tašna* 'briefcase' and *saher-torta* 'Sacher cake' *akten* and *saher* are not simple word forms in Serbian so they would be listed among unrecognized words. Quite a number of simple word forms found in this list belong to the compound proper names, like *Šri Lanka* 'Sri Lanka', *Skotland Jard* 'Scotland Yard', and *Ajfelova kula* 'Eiffel Tower'.

Specific patterns can be used in order to try to discover the compounds, as suggested in [15]. Useful patterns can be constructed by using the syntactic and semantic markers that are added to the entries in the dictionary of simple lemmas. For instance, all the adjectives that represent colors are marked in the Serbian dictionary of lemmas by the marker `+Col`, and thus the pattern `<A+Col>`

`<N>`[4] used on various texts can reveal quite a number of compounds. Among the retrieved compounds there are common names, such as *bela kafa* 'coffee with milk', *crno tržište* 'black market', *siva ekonomija* 'gray economy', but also quite a number of proper names: *Crno more* 'Black Sea', *Crveni krst* 'Red Cross', *Žuta reka* 'Yellow river'. Some, but not many, additional compounds are retrieved by the pattern `<A+Col> <A>* <N>`, for instance *siva moždana masa* 'cerebral cortical gray matter'.

Beside the color maker `+Col`, other markers that can be used in the same pattern are those indicating relational adjectives such as `+Zool`, `+Mat`, and `+NProp+Top`, referring to animals, substances and geographical proper names, respectively. They allow to retrieve compounds such as *labudji pev* 'swan song', *staklena bašta* (literally 'glass garden', meaning 'greenhouse'), *šećerna bolest* (literally 'sugar disease', meaning 'diabetes mellitus'), *Saudijska Arabija* 'Saudi Arabia', *Jadransko more* 'Adriatic Sea', *Versajski mir* 'the Peace Treaty of Versailles', *užička pršuta*, a type of prosciuto from Užice (town in Serbia), etc. A certain number of compounds is also retrieved with the marker `+Ord` that denotes ordinal numbers, e.g. *treći svet* 'third world', *na prvi pogled* 'at the first sight'.

Some more complex patterns were used to retrieve compound nouns. A grammar in a form of finite state graphs has been developed that recognizes functions, professions and titles of people. It is particularly successful when applied to newspaper texts in order to retrieve personal names followed or preceded by such designations ([10]). Some compounds retrieved are *narodni heroj* 'national hero', *književni kritičar* 'literary critic', *vršilac dužnosti* 'acting officer', *kandidat za predsednika* 'candidate for the president', etc.

4 Inflection

Morphological dictionaries of simple word forms of the DELAF type are produced automatically from the dictionaries of lemmas (of DELAS type). Namely, an inflectional class code is attached to every lemma which determines the FST that produces all the members of the lemma's paradigm with appropriate values of grammatical categories. The programming environments such as Intex[5], Unitex[6] and NooJ[7] incorporate these transducers and enable the automatic production of the DELAF. All three systems enable work with compounds but do not offer means for automatic production of a DELACF. In NooJ a step has been done towards it by introducing some new operators that can be used for inflection, for instance, "go to the end of the previous word", but serious linguistic problems have not been tackled (see [19]). Another, lexicographically based, approach relying on a systematic compound per compound description ([11]) is too specific to be efficiently applied to Serbian. In other corpus-oriented contexts the

[4] This is the over-simplified version of the pattern used; the actual pattern is more complex since it takes care about the agreement.
[5] http://msh.univ-fcomte.fr/intex/
[6] http://www-igm.univ-mlv.fr/~unitex/
[7] http://www.nooj4nlp.net

inflectional morphology of compounds is dealt with via automatic stemming or lemmatizing of their component words, or via combinations of all their inflected forms. As discussed in [16], these methods suffer from excessive generalizations or from overlooking of exceptions.

The Xerox finite-state lexicon compiler, *lexc* ([7]), based on the two-level morphology, allows the representation of inflectional and derivational morphology in terms of morpho-phonological phenomena. In particular, via a cascaded composition of lexical transducers, it enables the description of inflected forms of compounds. An example for French shows that a number of mechanisms, including unification, allows the *lexc* rules to combine different inflected forms of single constituents in order to obtain the inflected forms of the whole compound, as is the case in our formalism described below. The *lexc* rules probably allow to cover most of the compound inflection paradigms within the same framework as the simple words' morphology. However, if the description of the simple words has been done by a different formalism, its integration to *lexc* for compounds' inflection seems difficult. Moreover, it remains to be examined how some morpho-syntactic variants of compounds, which require constituent insertion, deletion, or order change, may be modeled by *lexc* rules.

The problem of the inflection of compounds is regarded as serious for English and French. However, in [17] the most complex example for English is *student union* that has three possible single forms: *student union*, *students union*, and *students' union*, and three possible plural forms: *student unions*, *students unions*, and *students' unions*. For Serbian, and other Slavic languages, the problem is more complex. For instance, in Serbian, nouns are characterized by four categories: gender, number, case, and animateness, and they inflect in two of them, number and case. There are seven cases and three numbers: singular, plural, and paukal that is used only with the small numbers two, three, and four, and only in genitive and accusative case. A compound noun, like a simple noun, has thus, in most cases, 16 different possible realizations, and in order to produce them different agreement conditions have to be taken into consideration for all of its *characteristic constituents* (CC), that is, the headword and all the constituents that agree with it. For instance, if the headword is a noun, and the other CC is an adjective, than the adjective has to agree with the noun in gender, which can change in a noun paradigm but not freely, in number and case for which the noun inflects, and in certain cases with the animateness which is fixed for a noun. In addition, the adjectives inflect in degree (positive, comparative, and superlative), and definiteness (definite and indefinite), independently from the noun.

In [16] a method is suggested that enables an effective inflection of compounds that satisfies both the condition of *correctness* and *exhaustivity*, that is, nothing that does not belong to the compound's paradigm is produced, and everything that belongs to it is. The method is based on a "two-level" approach[8] that separates the inflectional characteristics of compounds from the inflectional characteristics of its constituents. Namely, two compounds as a whole can behave in

[8] Not to be confused with Koskenniemi's two-level morphology.

the same way, although their characteristic constituents inflect in different ways, for instance, *Ujedinjene nacije* 'United Nations' and *Crno more* 'Black Sea'. As compounds they have the same structure, that is, the structure of an adjective followed by a noun, adjective and noun agree in gender, number, and case, and noun in either compound does not inflect in number. The constituent adjectives and nouns in the given examples inflect in a different way, as suggested by their different inflectional codes listed in the Serbian DELAS:

```
(ujedinjen,A1 nacija,N600) and (crn,A10 more,N300)
```

Moreover, the constituent noun is in the first compound always in plural, and consequently the compound has only plural number, while in the second compound the constituent noun is only in singular and as a result the compound is also always in singular. However, according to this method, these two compounds would belong to the same class, regardless of the different characteristics of their constituents.

In order to describe the inflectional characteristics of compounds, two formalisms are defined: *inheritance* and *unification*. The compound can inherit some category values from some of its constituents through the inheritance mechanism, for instance in the example of *Ujedinjene nacije* the value of the category number is inherited from the headword *nacije*, and it is plural. Some categories are neither fixed nor inherited but can take all the values allowed for them. These values, however, have to be in accord for the CC, which is established by the unification mechanism. For instance, for the same example, the different forms of the CC for the category case, when category number, gender and animateness are inherited, are as follows:

```
ujedinjene:1   nacije:1        ujedinjene:5   nacije:5
ujedinjenih:2  nacija:2        ujedinjenim:6  nacijama:6
ujedinjenim:3  nacijama:3      ujedinjenim:7  nacijama:7
ujedinjene:4   nacije:4
```

The word forms in these two columns cannot combine freely, only those that have the same value of the case category can combine. The unification mechanism is, thus, similar to the natural join operation in relational algebra.

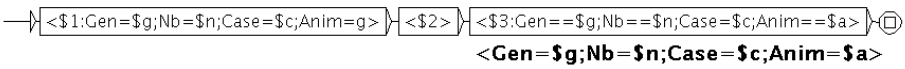

Fig. 1. The inflectional FST for the compounds of the type *Ujedinjene nacije*

These two mechanisms are supported by a new type of a graph[9], which generates all the inflected forms of a compound. Such a graph for compounds *Ujedinjene nacije* and *Crno More* is presented on Figure 1. All the compound

[9] These FSTs rely on Unitex inflectional FSTs.

constituents are represented in the FST by ordinal numbers, non-alphabetic characters being constituents on their own. The headword is the third constituent ($3) since the values of the categories gender (Gen), number (Nb), and animateness (Anim) are inherited from it (which is signaled by the double equal sign). The first ($1) and third constituent both inflect in case (signaled by the single equal sign for Case), but they have to agree (signaled by the use of the same variable $c for the category Case). The use of the same variable $c for Case in the first and the third constituent actually extends the only path in the graph in Figure 1 into seven paths with seven different output values – if two different variables were used that path would be extended into 49 paths. To continue the analogy with the relational algebra, that would correspond to the Cartesian product. Two DELAC entries for the given example illustrate the usage of the inflectional graph (named NC_A3XN2) from the Figure 1 and the "two-level" approach:

```
Ujedinjene(Ujedinjen.A1:aefp1g) nacije(nacija.N600:fp1q),NC_A3XN2
Crno(crn.A10:aens1g) more(more.N300:ns1q),NC_A3XN2
```

In order to use this method, the compounds have to be analyzed and classified according to their different characteristics:

1. *The number of constituents.* This is usually not difficult to establish, but this point is connected to the establishment of the lemma. Consider the adjective *vojno-tehnički* 'military and technical' that can also be written *vojnotehnički*; however, the latter cannot be chosen for lemma since it is not possible to unambiguously distinguish the constituents in it. For the constituents that do not inflect in the compound the corresponding DELAF entry need not be given. As a result, one compound inflectional class can contain syntactically different compounds. For instance, *Ministarstvo za informacije* 'Ministry for Information' and *Ministarstvo spoljašnjih poslova* 'Ministry of Foreign Affairs' would be in one inflectional class although the first one has the structure `<N> <PREP> <N>` and the second `<N> <A:2> <N:2>`, because in both cases the last two constituents do not inflect.
2. The identification of the constituents that can be omitted, e.g. *profesor engleskog jezika* 'professor of English language' is often used in a shorter form *professor engleskog* (the third constituent is optional).
3. The identification of optional replacements, e.g. *žiro račun* 'giro account' can also be written with the hyphen *žiro-račun*.
4. The identification of the allowed word reordering, e.g. *Božji sud* and *sud Božji* 'ordeal'.
5. The identification of characteristic constituents and their agreement conditions. Although this seems straightforward, it is by no means so. Consider the example of a compound adjective *gladan kao vuk* 'hungry as a wolf'. The characteristic constituent is the adjective *gladan* that inflects in gender, number, case, but can inflect neither in degree (**gladniji kao vuk* is not syntactically correct) nor animateness. The problem is whether *vuk* inflects as

well, and, if it does, how it agrees with the noun to which the adjective is applied. The following examples from the CCS illustrate the problem:

(a) *Posle takvih vežbi Grmalj je bio <gladan kao vuk>*. 'After these exercises Grmalj was hungry as a wolf'.
(b) *Kad dodju sa treninga, <gladni kao vukovi> i otvore frižidera,...* 'When they come back from training, hungry as wolfs and open the refrigerator...'
(c) *Ako ste <gladni kao vuk>, možete pojesti i porciju barenog žutog pirinča...* 'If you are hungry as a wolf you can eat a portion of boiled yellow rice...'
(d) *Posle pušenja kanabisa osoba je pospana, nervozna, <gladna kao vuk>*. 'After smoking cannabis, one is sleepy, nervous, hungry as a wolf.'

The examples (a) and (b) show that *vuk* inflects in number and agrees with the noun or pronoun the adjective is applied to. The example (c) shows that the adjective can be in plural and *vuk* in singular if the plural form is used as a form of a polite address. The example (d) shows that the adjective can be in a feminine form although *vuk* is in masculine (**gladna kao vučica*, 'hungry as a female volf' is not used). After this considerations, the inflectional graph for this type of compound adjectives is given on Figure 2.

6. The identification of the categories for which the constituents inflect and those for which the values are inherited. For instance, in *Crno more* the noun *more* does not inflect in number (**Crna mora*), in *redovni profesor* 'full-time professor' the adjective *redovan* does not inflect in degree (**redovniji profesor*) and only its definite forms are used (**redovan profesor*).
7. The identification of the output values of grammatical categories. For many types of compounds this is straightforward, for instance for the compounds with the structure `<A> <N>`, the compound will inherit its gender and animateness from the noun, it will inflect in number (or inhert the number) and in case. The following examples show that some compounds are more complex:

(a) *Komanda Unprofora za bivšu <Bosnu i Hercegovinu> nije prihvatila...* 'The command of Unprofor for the former Bosnia and Herzegovina has not accepted...'
(b) *<Bosna i Hercegovina> su na 70-tom mestu....* 'Bosnia and Herzegovina are on the 70th place...'.
(c) *<Kosovo i Metohija> je postalo leglo organizovanog kriminala...* 'Kosovo and Metohija has become the nest of the organized crime...'.
(d) *<Kosovo i Metohija> su bili, sada su i ostaće multietnička sredina.* 'Kosovo and Metohija were, are now and will remain a multiethnic'.

The examples (a) and (b) show that the gender of *Bosna i Hercegovina* 'Bosnia and Herzegovina' is feminine because both *Bosna* and *Hercegovina* have feminine gender. Its number, however, can be both singular (a) and plural (b). Even more complex is the case of *Kosovo i Metohija* 'Kosovo and Metohija'. If used as singular its gender is neuter since *Kosovo*, the first constituent is neuter (c), but if used as plural its gender is masculine (d), although neither *Kosovo* nor *Metohija* are.

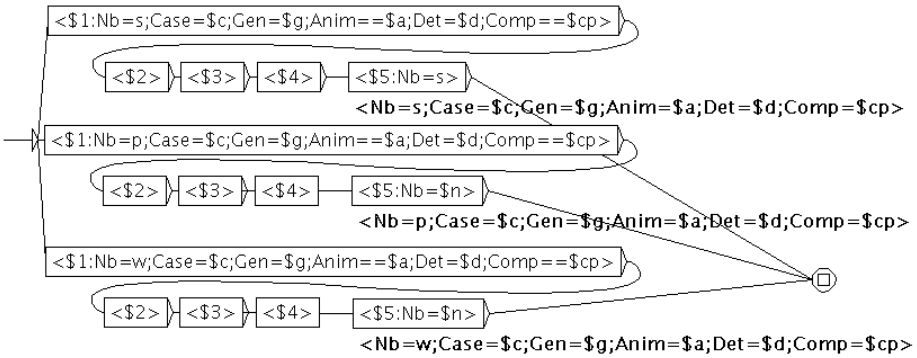

Fig. 2. The inflectional FST (called NC_A3XN2) for the compound adjective of the type *gladan kao vuk*

The application of this method to the Serbian compounds has shown that the "two-level" principle cannot be applied to all cases. Namely, in Serbian there are come classes of nouns that change their gender with the number: *papa.ms* 'pope' vs. *pape.fp*, *sudija.ms* 'judge' vs. *sudije.fp*. In this case, gender is not an independent category as it is usually treated, so it can be neither inherited nor can it inflect freely. As a consequence, when a noun of this type is a constituent of a compound a different FST for the compound inflection has to be constructed. In addition, there are both nouns and adjectives that do not inflect at all, and ask for a special treatment as well.

Although the principle of exhaustivity can always be satisfied, the principle of correctness is sometimes disrupted. Namely, in Serbian the adjectives for some cases and numbers have shorter and longer forms that are not treated as special categories in the traditional grammars, and thus we have not specifically marked them in the Serbian DELAS. In compounds, as well as in nominal phrases, these different forms cannot combine. However, since we have not marked them appropriately some erroneous compound forms are generated, as for *okružni javni tužilac* 'district attorney':

```
okružnoga javnoga tužioca,okružni javni tužilac.NC+Comp:ms2v
*okružnoga javnog tužioca,okružni javni tužilac.NC+Comp:ms2v
*okružnog javnoga tužioca,okružni javni tužilac.NC+Comp:ms2v
okružnog javnog tužioca,okružni javni tužilac.NC+Comp:ms2v
```

The application of the compound inflection FSTs has thus detected a serious flow in the dictionaries that we have to correct in order to achieve a full correctness. On the other hand, creating an extensive DELAC/DELACF sample for an inflectionally reach language such as Serbian allowed for the new compound inflection formalism and software to undergo their first large-scale test of adequateness and correctness.

5 Production

The final step in the production of the dictionary of compounds is the preparation of the list of entries in the desired format. Due to the "two-level" approach the preparation of one entry in DELAC is much more complex then the preparation of one entry in DELAS. Namely, besides the correct inflectional code of a compound, one has to add, for each constituent that inflects, the full DELAF entry of the form that appears in a compound lemma, that is: (a) simple word lemma; (b) inflectional code; (c) grammatical categories. In order to facilitate this work a module has been developed within the software named WS4LR — workstation for the lexical resources ([9]). First of all, this module enables the existing entries to be copied, and in that way for the compounds that share the same structure the compound inflectional code is copied. For each word form that inflects, the Unitex routines are invoked that retrieve from the appropriate DELAF dictionaries all the necessary information. Often, more then one DELAF entry satisfies the query, and in that case the user has to choose the correct one. The only case when the user actually has to fill in all the fields is when the word form does not appear in the dictionary of simple words. The only data that has to be entered for all the new entries are the semantic markers since, in general, they cannot be inherited from the constituent lemmas.

6 Conclusion and perspectives

The dictionary of compounds for Serbian has at this moment around one thousand lemmas. Much more of them have been collected but have not yet been classified and processed accordingly. However, now that all the prerequisites have been achieved it is expected that this dictionary will grow quickly. The description of compounds is not finished. Some lemma variations can be described by the mechanism shown in section 4, for instance for the lemma *ministar za saobraćaj* 'minister for traffic' the syntactically variant form *ministar sobraćaja* can be generated. However, for *kandidat za predsednika* 'candidate for the president' the variant form *predsednički kandidat* cannot be produced since its constituent is the relational adjective *predsednički* derived from *predsednik* and it is not part of the noun paradigm. Similarly, from many compound geographic names simple derivational forms are obtained, e.g. *Novi Sad*, town in Serbia, and *novosadski* 'related to Novi Sad', *Novosadjanin* 'the inhabitant of Novi Sad', and presently they have to be treated separately.

A reliable quantitative and qualitative evaluation of the proposed methodology will only be possible when the dictionary reaches a large-coverage size. However, having linguistically studied various, even rare, compound inflection paradigms for Serbian, Polish, English and French, allows us to believe that a very high percentage of existing compounds may be correctly described by our formalism. Naturally, this human-controlled process will be labor intensive but will also allow a very reliable and easily maintainable lexicographic data.

References

1. Corbett, G. G.: Number. Cambridge University Press (2000)
2. Courtois, B., Silberztein, M., eds.: Dictionnaires électroniques du français. Langue Française, 87. Larousse (1990)
3. Downing, P.: On the Creation and Use of English Compound Nouns. In Language, 153(4), Linguistic Society of America (1977)
4. Gross, G. Définition des noms composés dans un lexique-grammaire. In Langue Française, 87, Larousse, Paris (1990)
5. Habert, B., Jacquemin, Ch.: Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. In TAL, 2 (1993)
6. Jacquemin, Ch.: Spotting and Discovering Terms through Natural Language Processing. MIT Press (2001)
7. Karttunen, L.: Finite-State Lexicon Compiler. Technical Report. ISTL-NLTT2993-04-02. Xerox Palo Alto Research Center. Xerox Corporation (1993)
8. Koeva, S, Krstev, C., Obradović, I., Vitas,D.: Resources for Processing Bulgarian and Serbian — a brief overview of Completeness, Compatibility and Similarities. In S. Piperidis and E. Paskaleva, eds.: Workshop on Language and Speech Infrastructure for Information Access in the Balkanic Countries, 25 September 2005, Borovets, Bulgaria. (2005) 31–38
9. Krstev, C. Stanković, R., Vitas, D., Obradović, I.: WS4LR: A Workstation for Lexical Resources. In: Proc. of LREC'06, Genoa, ELRA (2006).
10. Krstev, C., Vitas, D., Gucul, S.: Recognition of Personal Names in Serbian Texts. In G. Angelova, ed.: Proc. of the International Conference Recent Advances in Natural Language Processing, 21-23 September 2005, Borovets, Bulgaria. (2005) 288–292
11. Kyriacopoulou, T., Mrabti, S., Yannacopoulou, A.: Le dictionnaire électronique des noms composés en grec moderne. In Lingvisticae Investigationes, 25(1), John Benjamins B.V. (2002) 7–28
12. Laporte, E.: Reduction of lexical ambiguity. Lingvisticae Investigationes, 24(1), John Benjamins B.V. (2001) 67–103
13. Mikheev, A., Grover, C., Moens, M: Description of the LTG System Used for MUC-7. In Proceedings of the 7th Message Understanding Conference (MUC-7).
14. Monachini, M., Soria, C.: Building Multilingual Terminological Lexicon for Less Widely Available Languages. In Proc. of LTC'05, Poznań, Poland (2005) 129–133
15. Ranchhod, E.M.: Using Corpora to Increase Portuguese MWE Dictionaries. Tagging MWE in a Portuguese Corpus. Proc. of the Corpus Linguistics Conference Series 1(1) (2005) [to appear].
16. Savary, A.: A formalism for the computational morphology of multi-word units. Archives of Control Sciences, 15(LI) (2005) 437–449
17. Savary, A.: Multiflex — User's Manual and Technical Documentation, version 1.0. Technical Report 285, LI-University of Tours, Tours (2005)
18. Silberztein, M.: Le dictionnaire électronique des mots composés. Langue Française, 87 (1990) 71–83
19. Silberztein, M.: NooJ Manual. Université de Franche-Comté (2005) http://perso.wanadoo.fr/rosavram/NooJ
20. Vitas, D., Krstev, C.: Derivational Morphology in an E-Dictionary of Serbian. In Z. Vetulani, ed.: Proc. of LTC'05, Poznań, Poland (2005) 139–143
21. Vitas, D., Pavlović-Lažetić, G., Krstev, C., Popović, Lj., Obradović, I.: Processing Serbian Written Texts: An Overview of Resources and Basic Tools. In S. Piperidis and V. Karkaletisis, eds.: Workshop on Balkan Language Resources and Tools, 21 November 2003, Thessaloniki, Greece. (2003) 97–104