An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds

Cvetana Krstev, Duško Vitas

Abstract. In this paper we present the process of creating a comprehensive morphological dictionary of compounds for Serbian. This dictionary should be compatible with existing large morphological dictionaries of simple words for Serbian. Due to the complexity of Serbian morphology, the production of a dictionary of compounds is not an easy task. In this paper we present a procedure that automatically produces lemmas for such a dictionary for a given list of compounds. In making decisions this procedure relies on data found in the e-dictionaries of simple words. We evaluate the procedure developed on several different sets of data.

Keywords: electronic dictionary, Serbian language, morphology, inflection, compounds

1 Introduction

We have been developing a morphological electronic dictionary of Serbian for many years now. Our e-dictionaries follow the methodology and format (known as DELAS/DELAF) presented in (C. Courtois et M. Silberztein 1990). The size of our e-dictionaries of simple forms is considerable: they have today more then 121,000 entries (C. Krstev 2008).

In recent years the interest in multi-word units and compounds has been growing rapidly, and they have been analyzed from various points of view (Ch. Jacquemin 2001). They have roused the interest of the Serbian NLP community as well, e.g. some initial work has been done on automatic terminology extraction (G. Nenadić and I. Spasić. 2008). However, our interest is mainly in the morphological description of compounds that would be compatible with methodology used for simple words. We have chosen the most suitable frame for this purpose (described in (A. Savary 2005)) which relies on the usage of Finite-State Technology. The final aim is to produce a counterpart of DELAS/DELAF for compounds – DELAC/DELACF. The content of these dictionaries and the problems in their development can be illustrated with one example: for the given compound *drveni duvački instrument* 'woodwind instrument' we would like to enter the following entry into the dictionary of compounds DELAC (C. Krstev, D. Vitas and A. Savary 2006):

drveni(drven.A6:adms1g) duvački(duvački.A2:adms1g)
instrument(instrument.N29:ms1q), NC_AXAXN+Conc

The information between the parentheses follows each compound component that inflects in a compound. It describes how each particular component inflects and which of its inflected forms occurs in a compound lemma. A compound inflectional code that follows a comma sign determines which inflectional forms are used in a compound and how they agree with each other. All the information found in this entry shall allow for the automatic production of all inflected forms for the dictionary DELACF, e.g. the inflected form for the singular instrumental case would be

drvenim duvačkim instrumentom,drveni duvački instrument.NC_AXAXN:ms6q

The production of a lemma for the DELAC dictionary proceeds in several steps:

1. Each component determines its lemma in the DELAS dictionary together with its inflectional class code, and its grammatical categories from the DELAF dictionary. For instance, for *drveni* the lemma is *drven*, its inflectional class code is A6, and the grammatical categories of the form *drveni* are :adms1g;
2. The inflectional class code for a given compound (e.g. NC_AXAXN) is determined;
3. The syntactic and semantic markers for a given compound (e.g. +Conc) are determined.

Performing all these steps manually is unacceptable. It is shown in (A. Savary 2008, 40-41) that among the eleven analyzed tools for multi-word inflection description only FASTR (Ch. Jacquemin 2001) supports some kind of automated multi-word lexicon creation. In order to facilitate this task, we have developed a special tool, a WorkStation for Lexical Resources - WS4LR (C. Krstev, R. Stanković, D. Vitas and I. Obradović 2008) that helps in obtaining some of the information necessary from already developed dictionaries of the type DELAS/DELAF and thus reduces the number of errors in DELAC entries. The development of a DELAC dictionary for Serbian nevertheless remains slow. Due to this, we have decided to develop a procedure for fully (or almost) automatic construction of a DELAC type dictionary from the given list of compounds that would be compatible with the chosen formalism (A. Savary 2005).

2 A Short Description of Test Data

Our set of test data consists of our present dictionary of compounds which has 3,050 lemmas covering different parts of speech. Among them are 2,650 nouns and adjectives and, since they

only inflect, they are assigned an inflectional class code. Each inflectional class is associated with one inflectional transducer (as described in (A. Savary 2005)) which controls the production of all inflected forms. The nouns presented now in our dictionary of compounds are covered by 70 different transducers, while there are 14 transducers for the adjectives. There are less than 70 different noun compound structures since some inflectional transducers represent variations of the basic structure, as represented in the following table. The same is true for the adjectives.

Inflectional	**Example**	**Translation**	**Explanation**
NC_NXN	lekar akušer	obstetrician	Both components inflect and agree in the case and number; a compound inherits grammatical categories from the first component.
NC_NXNF	kit ubica	killer whale	the gender of the second component changes in plural from masculine to feminine
NC_NXN3	Kamen-temeljac	foundation stone	the separating hyphen can be replaced by a space
NC_NXN2	Kongo Brazavil	Congo-Brazzaville	neither component inflects in number
NC_NXN2m	Kneževina Monako	The Principality of Monaco	the second component may inflect, but may also remain uninflected

In order to formulate our strategy for the detection of the structure and inflectional properties of compound lemmas we synthesized our test data set and collected some numerical information. First, we established what grammatical information corresponds to the components of compounds belonging to a particular inflectional class (see the following table). We found, for instance, that the inflectional class NC_NXN3 applies to four different combinations of grammatical categories of the first and second component, that the first and the second component always agree in gender (according to our test data) but need not agree in animatness (line 3 in our table; *država* 'state' is inanimate while *članica* 'woman member' is animate). The class NC_NXN2m is associated with only one combination of grammatical categories which is characterized by the fact that the components differ in gender.

Inflection	Gramm.Categories	Frequency	Example	Translation
NC_NXN2m	_:fs1q_:ms1q	22	*Republika Kipar*	The Republic of Cyprus
NC_NXN3	_:fs1q_:fs1q	2	*zvezda-padalica*	falling star
	_:fs1q_:fs1v	2	*država-članica*	member state
	_:ms1q_:ms1q	2	*kamen-temeljac*	foundation stone
	_:ms1v_:ms1v	2	*golub-pismonoša*	carrier pigeon

This type of data is not the only type relevant. We have also derived data about the combinations of grammatical categories and inflectional classes that are associated with them. The first three lines in the following table show that the same combination of grammatical categories of an adjective and a noun can be found in compounds described by three different compound inflectional classes. We have found that the combination where the first component is in the feminine gender and the second component is in the masculine gender is associated with only one inflectional class — NC_NXN2m (line 4 in the next table). This information is not conclusive. For compounds of the type X+N or N+X where 'X' stands for a word form that does not inflect in a compound, we cannot deduce from our DELACF what the grammatical categories of the component 'X' might be since they are not given in compound lemmas. This is generally the problem with the type of compounds whose components are both genuine nouns in the nominative case. In some cases both inflect (the first component is the head) while in others only the second component inflects (the second component is the head). This makes the distinction between N+N, X+N and N+X difficult to express. Examples are *zvezda vodilja* 'guiding star' that has the structure N+N and the inflectional class code NC_NXN, *kristal šećer* 'granulated sugar' that has the structure X+N and the inflectional class code NC_2XN, and *diler deviza* 'foreign currency dealer' that has the structure N+X and the inflectional class code NC_N2X. Here all six compound components are Serbian nouns in the nominative case singular (with other possible interpretations).

Grammatical Categories	Inflectional	Frequencies
:aefs1g:fs1q	NC_AXN	352
:aefs1g:fs1q	NC_AXN3	112
:aefs1g:fs1q	NC_AXNr	3
_:fs1q_:ms1q	NC_NXN2m	22

3 The Description and Implementation of the Strategy

On the basis of our test data and the analysis performed, we have designed the strategy which is used by our procedure for the construction of a complete compound lemma for the dictionary of the DELAC type. The strategy and the procedure are independent, that is, in general the changes in the strategy do not inflect the procedure itself. Such a design of our system makes it possible to experiment with various strategies. The strategy itself consists of a list of rules where each rule defines the conditions that should be satisfied by the components of a particular compound and/or

separators between them in order to assign a certain inflectional class to it. The rules are ordered, which means that the rules listed first are applied first. All the rules are recorded using XML which makes them easy to understand and to manipulate using a standard XML tool. The conditions defined for each rule are of two types: 1) conditions of the first type specify grammatical categories of compound components that should be satisfied and they usually apply to those components that inflect (`<RuleGenCond>`); 2) the conditions of the second type, however, specify additional conditions like semantic and/or syntactic markers that should be satisfied (`<RuleSpecCond>`). This can best be illustrated by the example of one rule:

```
<Rule ID="25" CFLX="NC_N4X" CflxGroup="NC_N4X" >
      <RuleGenCond>
            <Word ID="1" POS="N" Flex="true" Case="1" Num="s" Cond="!$FLXN"/>
            <Word ID="2" POS="MOT" Flex="false"  />
            <Word ID="3" POS="MOT" Flex="false" />
      </RuleGenCond>
      <RuleSpecCond ID="1" Example=”bolest ludih krava/ mad cow disease”>
<!-- The second and the third component are in the genitive case -->
            <Word ID="1" />
            <Word ID="2" POS="A,N,PRO" Case="2" />
            <Word ID="3" POS="N" Case="2"/>
      </RuleSpecCond>
      <RuleSpecCond ID="4" Example="raketa zemlja-vazduh/surface-to-air
missile"> <!-- A separator between a second a the third word is a hyphen -->
            <Word ID="1" POS="N" Case="1" Num="s" />
            <Word ID="2" POS="MOT" Sep="-"/>
            <Word ID="3" POS="N" Case="2"/>
      </RuleSpecCond>
</Rule>
```

This rule applies as follows: if the first component satisfies (according to the dictionary of simple words) the grammatical conditions (which imply that the first component has to be a noun in the singular as well as in the nominative case), and if the second and the third component and separator between them satisfy one of the remaining additional conditions, then the rule class NC_N4X will be suggested for the given compound. The first additional condition is satisfied if the second compound component is an adjective, a noun or a pronoun in the genitive case and if the third compound component is a noun also in the genitive case. The example given above is *bolest ludih krava* ‘mad cow disease’ where *ludih krava* is a genitive case plural form of the phrase *luda krava* ‘mad cow’. In order to keep the example short we have not listed all the additional conditions. The next example illustrates another rule feature. Namely, in accordance with the formalism used in a Multiflex system (A. Savary 2005; Paumier 2008), the general condition states that whatever the gender and animatness of the third component (a noun) are, the

first and the second component (adjectives) have to agree with them. For instance, if the value of the category ‘Gen’ (gender) for a component is ‘=$g’ it suggests that it can be any value allowed for this category, and if it is ‘$g’ than the value of this category has to agree with that of another component that established the value of $g.

```
<Rule ID="26" CFLX="NC_AXAXN1" CflxGroup="NC_AXAXN" >
      <RuleGenCond>
            <Word ID="1" POS="A" Flex="true" Case="1" Gen=”$g” Anim="$a"/>
            <Word ID="2" POS="A" Flex="true" Case="1" Gen=”$g” Anim="$a" />
            <Word ID="3" POS="N" Flex="true" Case="1" Gen=”=$g” Anim="=$a" />
      </RuleGenCond>
      <RuleSpecCond ID="1" Example="Prvi svetski rat/First world war">
<!—All three components are in singular and the first component is written
with upper case -->
            <Word ID="1" Num="s" Cond="$SWUC" /> <Word ID="2" Num="s" />
            <Word ID="3" Num="s" />
      </RuleSpecCond>
      <RuleSpecCond ID="2" Example=”gornji disajni putevi/upper respiratory
tract>      <!—All three components are in plural -->
            <Word ID="1" Num="p" /> <Word ID="2" Num="p"/>
            <Word ID="3" Num="p" />
      </RuleSpecCond>
</Rule>
```

The general condition is satisfied if the first two components are adjectives and the third component is a noun, if all three components are in the nominative case, and if the gender and the animacy category of the first two components agree with the corresponding categories of the third component. The first additional condition is satisfied if all three components are in the plural and the first one starts with a capital letter.

For some inflectional classes we have more then one rule. These rules are modeled on different conditions and they have a different order in their strategy that reflects the probability of their application, which is calculated on the basis of the frequency data in our test sample. Also, for some compounds more than one solution is offered due to problems explained at the end of the last section; it is the responsibility of the order of rules to offer the most probable solutions first.

All conditions, both grammatical and additional, are checked against our morphological e-dictionaries of simple words. The dictionary look-up is performed by Unitex procedures (Paumier 2008) which we use in our own application WS4LR. A module for strategy development is integrated in this tool. The strategy itself can be used in two ways: it can be applied to a particular compound for which a DELAC entry has to be produced, or to a list of

compound candidates. In the latter case, the result can be inspected and corrected using MS Excel and the whole list can be integrated after inspection in the DELAC dictionary.

We have evaluated our procedure on four different test data. Obviously, the first is our existing DELAC dictionary. The results show that for almost 80% of compound nouns and adjectives the correct inflectional class was suggested by our strategy. The structure of a compound was correctly detected (but the inflectional transducers offered were not always correct) for more than 90% of compounds. Moreover, almost 90% of all these correctly or conditionally detected compound structures were offered as a first solution. Our strategy failed mostly in cases in which some compound component was not in the e-dictionaries of simple words, for instance, in *haus-majstor* 'building manager' the first component *haus* is not used as a simple word.

The first set of evaluation data consisted of a list of 209 compound Serbian proper names for inhabited places. More then 55% were correctly processed (60% compounds with the correct structure). The main reason for the poor results of our procedure on this data was the large number of unknown words, for instance, *Mala Moštanica* 'Small Moštanica' where *Moštanica* is neither among simple proper names nor common nouns. The second set of evaluation data consisted of a small part of the official list of compound names of professions (333 names). More then 85% of these names were correctly processed, while for most of the others our strategy did not offer any solution. Namely, some names of professions are quite artificially chosen and were thus too long to fit in the usual compound structure. The longest name in our subset had 12 components: *poslužilac uređaja za proizvodnju i preradu aditiva, začina, čaja, kafe i kavovina* 'attendant of the device for the production of additives, spices, tea, coffee and coffee based products'. The third evaluation data set consisted of data we have found in a log file of the search engine of a Serbian economic journal. We have selected 728 potential compounds from it. More then 84% were correctly processed (86% compounds had correct structure). The queries are usually free phrases with various structures and not compounds, and, in this case, this was the reason for most of failures. One example is *mala i srednja preduzeća* 'small and medium firms' whose structure A+X+A+N had not yet been detected among the compounds.

4 Conclusions

We feel that the results obtained are promising despite the fact that 100% accuracy cannot be achieved. We expect that with this procedure our dictionary of compounds will increase quickly. Besides this, it will reduce errors introduced by the human factor. Our procedure is not language dependent and it can be used by any language for which e-dictionaries of simple words exist in the LADL format. We envisage further development of our procedure. We plan to use our procedure for the inflection of free phrases submitted to web search engines. For such application, we would like to allow compound components to be compounds themselves. This would keep the number of possible structures low. For instance, in order to process the phrase *prošlogodišnji predsednički izbori u Srbiji* 'last-years presidential elections in Serbia' we would have to develop a new inflectional transducer for the structure **A+A+N+X+X**. However, *predsednički izbori* 'presidential elections' is a compound noun already in DELACF which reduces the above structure to the known structure **A+N+X+X**.

Bibliography

Courtois, B. & M. Silberztein (Eds.). 1990. *Dictionnaires électroniques du français*. Langue française. 87, Paris : Larousse.

Krstev, C. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionary*, Faculty of Philology, University of Belgrade, Belgrade.

Jacquemin, Ch. 2001. *Spotting and Discovering Terms through Natural Language Processing*, MIT Press.

Nenadić, G. & I. Spasić. 2008. "Towards Automatic Terminology Recognition in Serbian". In: *Formal Description of Slavic Languages - FDSL 2003*, eds. G. Zybatow, L. Szucsich, U. Junghanns and R. Meyer. Frankfurt am Main : Peter Lang, 3-17.

Savary, A. 2005. "Towards a Formalism for the Computational Morphology of Multi-Word Units", in: *Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland*, ed. Zygmunt Vetulani, 305-309.

Savary, A. 2008. "Computational Inflection of Multi-Word Units – A Contrastive Study of Lexical Approach", In: Linguistic Issues in Language Technologies, Vol. 1, No. 2, CSLI Publications.

Krstev, C., Vitas, D. and A. Savary. 2006. "Prerequisites for a Comprehensive Dictionary of Serbian Compounds". In: *FinTAL 2006*, *LNAI, vol. 4139*, eds. Salakosi, T., F. Ginter, S. Pyysalo and T. Pahikkala. Heidelberg : Springer, 552–564.

Krstev, C., R. Stanković, Vitas, D. and I. Obradović. 2008. "The Usage of Various Lexical Resources and Tools to Improve the Performance of Web Search Engines". In: *6th LREC International Conference on Language Resources and Evaluation, Marrakech, Marocco*.

Paumier, S. 2008. *Unitex 2.1 User Manual,* http://www-igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf.