Restructuring Lemmas in a Dictionary of Serbian

Cvetana Krstev*, Duško Vitas†

*Faculty of Philology, University of Belgrade
Studentski trg 3, CS - 11000 Belgrade
cvetana@matf.bg.ac.yu

†Faculty of Mathematics, University of Belgrade
Studentski trg 16, CS - 11000 Belgrade
vitas@matf.bg.ac.yu

Abstract

Traditionally produced lexical resources for Serbo-Croatian are not suitable for automatic processing of contemporary Serbian. More specifically, the processes of structural derivation, although very productive in Serbian, are not presented in either monolingual or bilingual dictionaries in a systematic way. The morphological e-dictionary of Serbian was initially produced on the basis of traditional resources and as such reproduces the same flaws. In order to overcome them two solutions are possible. One is to put in the e-dictionary all the lemmas produced by structural derivation, no matter whether they are recorded in traditional dictionaries and confirmed in the corpus of contemporary Serbian, and then to assemble them explicitly in a complex lemma by an appropriate lexical graph. This approach, however, implies an overproduction of lemmas and the construction of such a graph for each complex lemma. Another approach is to extrapolate the missing lemmas using morphological grammars that model specific morphological processes and that are applied only to the entries already in the dictionary, which is checked by using the lexical constraints. It is demonstrated that such morphological grammars enable a precise classification of processes of structural derivation in a way similar to the classification of inflective phenomena.

1. Introduction

Traditionally produced Serbian lexical resources are not suitable for automatic processing of contemporary Serbian. The reasons derive from the lack of a clear theoretical and/or methodological frame used for the lexicographic processing that leads to the imprecision of the grammatical description or unsystematic choice of lemmas.

The only complete explanatory dictionary of Serbo-Croatian (RMSMH, 1967) is based on manual and partial excerption from the Serbo-Croatian sources compiled before 1960. The consequences can be seen in the years-long postponement of finalization of the one-volume dictionary of contemporary Serbian that was envisaged as the abridgement of (RMSMH, 1967). Another, less visible consequence of the traditional frame of lexicographic processing is seen in the limited production of bilingual dictionaries.

In the field of the Serbian computational lexicography, which is being developed mostly outside the traditional lexicographic circles, resources have been produced that enable a different organization of the dictionary entry in Serbian. In this paper a part of this organization will be presented.

	LEMMA	GRAMMATICAL DEFINITION	DEFINITIONS	FREQUENCIES ON *SRPCOR*
1	*profesor*	m	title of teacher ... and person that has that title	2197
2	*profesorica*= *profesorka*	f	1. woman professor 2. wife of professor (**profesorova** žena)	*-ica* = 13 *-ka* = 212
3	*profesorski*	-a,-o	1. that is related to professors 2. that is characteristic for professors	67
4	*profesorčić*	m	dim. and pej. of professor	1

Table 1: Lemma *profesor* and its derivatives in (RMSMH, 1967)

2. The outline of the problem

We will illustrate the problems that traditional processing poses to the development of lexical resources with one characteristic example, the entry *profesor* 'professor' (see Table 1). It belongs to the class of masculine gender nouns with unmarked inflection, and as such represents the prototype of inflection for this class. Thus, in (RMSMH, 1967) only the information on gender is given in the part for the grammatical definitions, while all other information is being omitted, as for instance the information on animateness, that determines its accusative form.

Three entries that follow in (RMSMH, 1967) (rows 2-4 in Table 1) represent examples of structural derivation in Serbian. By structural derivation we mean the derivational processes that yield a predictable meaning of the word derived from the source word.

Structural derivation is a very productive phenomenon in Serbian that generates a number of lemmas, either by processes illustrated in Tables 1 and 3 or some other. Lemmas *profesorica* and *profesorka* are combined into one entry. They represent the phenomenon of gender motion that is traditionally seen as the process Nmasc → Nfem, where Nfem often has the meaning 'woman that is Nmasc'. In the part for grammatical definitions only their gender is given, which misleads to the conclusion that in this case the nouns also have an unmarked inflection. This, however, is not the case (Table 2).

	Unmarked inflection (N601)	*profesorica* (N651)	*profesorka* (N661)
nom.sing.	*žen-a 'woman'*	*profesoric-a*	*profesork-a*
voc.sing.	*žen-o*	*profesoric-e*	*profesork-a*
gen.pl.	*žen-a*	*profesoric-a*	*profesork-i*

Table 2: Inflectional properties of some feminine gender nouns

	Lemma	Grammatical definition	Definition	Frequencies on *SrpCor*
5	*profesorski*	Adv	in the way of professor	2
6	*profesorov*	AdjPoss	belonging to professor	20
7	*profesoričin*	AdjPoss	belonging to (woman) professor	0
8	*profesorkin*	AdjPoss	belonging to (woman) professor	2
9	*profesorčina*	f (Gen = m)	augmentative of professor	0

Table 3: Some derivatives of lemma *profesor* not recorded in (RMSMH, 1967)

Slovenian		Serbo-Croatian
Lemma	Grammatical definition	Translation
profesor	-ja, m	*profesor* 'professor'
profesorica	f	1. *profesorica* 'woman professor' 2. *profesorka*, profesorova žena
profesorski	-a,-o	*profesorski* 'as a professor'

Table 4: Lemma *profesor* and its derivatives in (Jurančić, 1981)

(Serbo-)Croatian		English
Lemma	Grammatical definition	Translation
profesor	m	*teacher, professor*
profesorica	f	*(women[30]) professor, professoress*
profesorka	f	see[16] *profesorica*
profesorski	adj	*teacher's, professor's, professorial*

Table 5: Lemma *profesor* and its derivatives in (Bujas, 2001)

English	sense	Croatian or Serbian
professor	1.	čovjek koji otvoreno iznosi svoje uvjerenje *'man that overtly presents his believes'*
	2.	(sveučilišni) profesor *'(university) professor'*
	3.	US u značenju nastavnik *'US, meaning teacher'*

Table 6: Lemma *profesor* and its derivatives in (Filipović, 1989)

Similar is true for the lemma *profesorčić*, diminutive of *profesor* (row 4, Table 1). This lemma is produced by structural derivation from the masculine gender noun. It should be noted, however, that the attribute 'pejorative' is not necessarily linked to diminutive and is context dependent. Lemma *profesorski,-a,-o (*row 3, Table 1) is the relational adjective: the fact is expressed by addition of endings -a,-o. This type of adjectives is derived regularly from the source noun, and in (RMSMH, 1967) they have prototypical definitions. The last column in Tables 1 and 3 give the frequencies of these dictionary entries established on the corpus of contemporary Serbian *SrpCor* (Vitas et al., 2003).

Lemmas from Table 3 are not recorded in (RMSMH, 1967) although they are derived by similar derivational processes. The definitions in Table 3 are prototypical and are used for other lemmas in the same dictionary. For instance, for *glumac* 'actor' some other derivational lemmas are listed, such as *glumičin* 'belonging to actress' and *glumčina* 'augmentative of actor', while some are missing, though recorded in SrpCor, such as *glumičica* 'diminutive of actress'. These lemmas form the part of a family of words with the noun *profesor* in a core, while relations between the lemmas in the family have a predictable meaning. For instance, notice in Table 1 that the possessive adjective *profesorov* is used in the definition of the meaning 2.2, *profesorova žena*, although the adjective itself is not an entry in (RMSMH,1967).

The consequences of the unsystematic processing of the structural derivation are reflected in the bilingual dictionaries as well. For instance, in the Slovenian/Serbo-Croatian dictionary (Jurančić, 1981) lemmas listed in Table 4 are recorded.

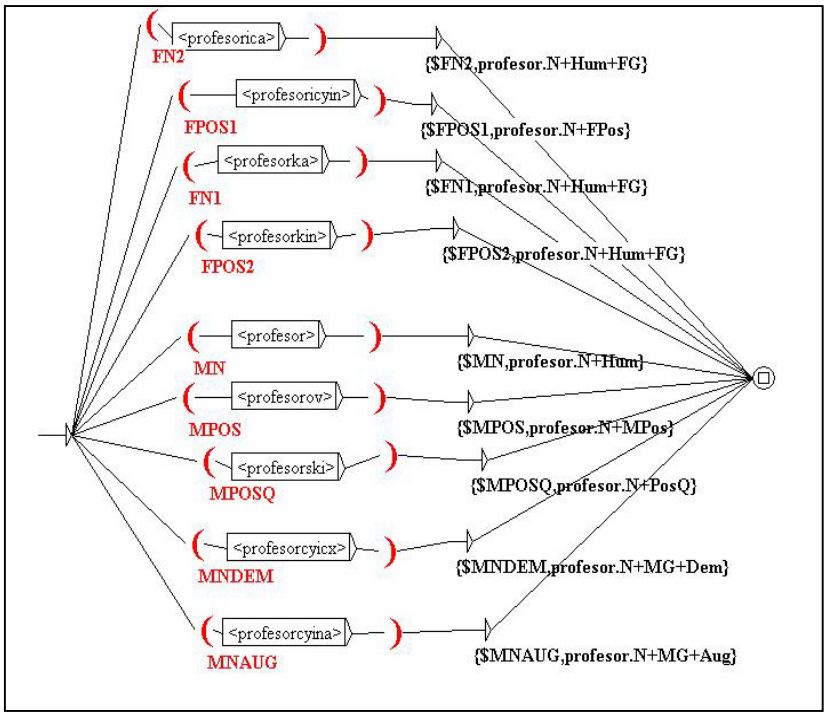

Figure 1: Graph for super-lemma *profesor*

Lemmas on the Serbo-Croatian side of this dictionary correspond to lemmas in (RMSMH, 1967) listed as 1, 2, and 3 in Table 1. The same entries are given also in the Croatian/English dictionary (Bujas, 2001), cited in a simplified way in Table 5. Superscripts in the English column indicate the gender motion (30), and reference from the variant prevailing in Serbian (16). Finally, only one entry is found in the English/Croatian or Serbian dictionary (Filipović, 1989) (Table 6).

It should be noted that indications of the inflective properties of the Serbo/Croatian equivalents in bilingual dictionaries are usually omitted.

Comparison of these data leads to the conclusion that the main meaning of lemma *profesor* in Serbo-Croatian is equivalent to lemma 'professor' in English. This main meaning is dispersed by structural derivation to several lemmas that are related to each other by prototypical definitions.

3. Morphological dictionary of Serbian and structural derivation

Morphological dictionary of Serbo-Croatian is being developed in the format that has been accepted in the frame of the network RELEX and it is used both for tagging, disambiguation and lemmatization of texts and as a preprocessor of text for the production of the tagged corpus of Serbian using system Intex (Silberztein, 1993). As to the initial choice of lemmas, the morphological dictionary of Serbian has reproduced those from traditional dictionaries. A class code that precisely describes its inflectional properties by means of a regular expression has been associated to each of these lemmas. For instance, class N1 is associated to

the noun *profesor*, and class N661 to the noun *profesorka*. The inflectional code N661 distinguishes lemma *profesorka* from, for instance, lemma *slika* 'picture', associated with code N612, by specifying the following features: there is no sibilarization in dative singular, the ending for the genitive plural is -i- instead of -a-, and for the vocative singular -a- instead of -o-.

The problem outlined in the section 2 poses a more complex question to the construction of the e-dictionary of Serbian (Vitas et al., 2001): Is it possible to assemble the results of the structural derivation in a complex lemma in a way that is similar to the description of the inflective characteristics, so that both in monolingual and bilingual Serbian dictionaries only one entry (e.g. 'professor') is listed with assigned codes that describe its derivational capability? From the point of view of the construction of the e-dictionary of simple lemmas, this problem can be solved in two ways. The first solution implies the construction of a complex lexical graph that would gather all the lemmas listed in Tables 1 and 2. Such a lexical graph for the super-lemma *profesor* is represented in Figure 1.

In this lexical graph every string in the square brackets refers to the corresponding entry in the e-dictionary (Vitas et al., 2000). This method enables the reduction of all these lemmas to one canonic form and the establishment of appropriate relations between them. The application of this graph to lemmatization gives the results presented in Table 7.

NATO-a na Jugoslaviju, berlinski	**{profesori,profesor.N+Hum}**	posebno naglašavaju
e boje neprijatnog znamenja ili -	**{profesorke,profesor.N+Hum+FG}**	ruskog jezika.
ma i rasklimanim klupama, niskim	**{profesorskim,profesor.N+PosQ}**	platama, razuz
nove u oblasti razmene studenata,	**{profesorskog,profesor.N+PosQ}**	kadra, naučni
tek sada svestan važnosti	**{profesorskog,profesor.N+PosQ}**	zvanja i svih
ku zahvalnost dugujemo pijanisti,	**{profesoru,profesor.N+Hum}**	Dušanu Trbojeviću

Table 7: Lemmatization with the graph from Figure 1

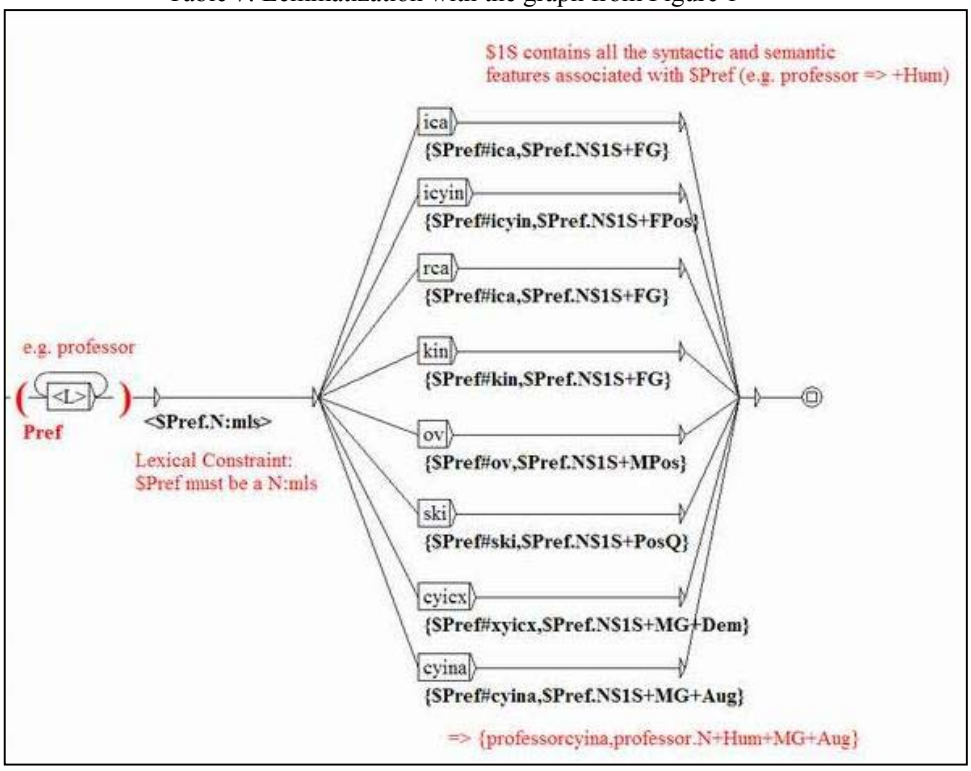

Figure 2: Graph of derivational class containing the lemma profesor

"kooperativni klxuč", zatim	**"kraljev**	čaj" i slično a kao najzanimljiviji,
salom, evo i izjava nekolicine	**Partizanovih**	fudbalera koji su na pomenutom jučerašnjoj
pet godina , za šankom je bio	**profesorov**	sin Branislav . Pošto je između njega
kilogram! Pa zato se i zove	**slonovo**	uvo! Od te paprike spravlja se ajvar, nacionalno

Table 8: Concordances produced by graph from Figure 2

This approach has several drawbacks. First, some of these lemmas (*profesorčina* and *profesoričin*), though possible, are neither listed in (RMSMH, 1967) nor have appeared in the corpus of Serbian. However, they have to be presented in the e-dictionary in order to be recognized by this lexical graph. Second, this graph is linked to the lemma 'professor', and not to the lexical class with its properties, and third, the lemma *profesor* is by no means an exception — the same properties apply to a large class of animate masculine nouns. Finally, as stated before the phenomena of structural derivation that are applicable to lemma *profesor* are not the only productive phenomena of this kind. Thus, another solution that makes use of the morphological grammars is being explored.

4. Morphological grammars

We have added a component in INTEX 4.30 that provides a different approach to the recognition of derived forms, by using morphological graphs that produce lexical constraints. (Silberztein, 2003). The basic idea is to extrapolate the missing lemmas by transducers that model certain morphological processes and that are applied only to the entries already in the dictionary. Applied to the example we used this means that only one lemma, *profesor*, needs to be present in the morphological e-dictionary. The remaining lemmas recorded in traditional dictionaries, as well as those that are not recorded, can be described by transducers that represent a type of a formal morphological grammar.

For instance, Figure 2 represents the morphological graph that describes the structural derivation processes, such as derivation of possessive adjective, gender motion, diminutives, augmentatives, and relational adjectives from the masculine gender nouns. In the case of the possessive adjective, the noun it is derived from is assigned during the text processing as a lemma to recognized adjective forms and a syntactic property +Pos is added to it. The application of this graph enables searching with the pattern <N+Pos>, that retrieves from the text all the possessive adjectives derived from the nouns in the dictionary, which themselves need not be in a dictionary. Several lines from the concordances produced are given in Table 8. It should be noted here that this solution enables the recognition of all possessive adjectives derived from the nouns presented in e-dictionary, and not only the particular cases, such as *profesorov*.

5. Conclusion

We have shown in this paper that lemmas derived by structural derivation in Serbian can be recognized in text without being explicitly recorded in a dictionary. As a consequence it is possible to reorganize the complete inventory of lemmas in dictionaries of Serbian, both monolingual and multilingual. Finite state transducers enable a precise classification of processes of structural derivation in a way similar to the classification of inflective phenomena.

6. References

RMSMH (1967). *Rečnik srpskohrvatskoga književnog jezika*, vol. 1-6, Beograd-Zagreb: Matica Srpska, Matica Hrvatska

Vitas, D., Krstev, C. Obradović, I., Popović, Lj. Pavlović-Lažetić, G. (2003): An Overview of Resources and Basic Tools for Processing of Serbian Written Texts, in Vangelis Karkaletsis and Stelios Piperidis, editors, *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece, (http://www.iit.demokritos.gr/skel/bci03_workshop)

Jurančić,, Janko (1981): Slovensko srpskohrvatski rečnik, Državna založba, Ljubljana

Bujas, Željko (2001): Veliki hrvatsko-engleski rječnik, Nakladni zavod Globus, Zagreb

Filipović, R. (1989): English-Croatian or Serbian Dictionary, Školska knjiga, Zagreb

Silberztein, M.D. 1993. *Le dictionnaire électronique et analyse automatique de textes: Le systeme INTEX*, Paris: Masson

Vitas, D., Krstev, C. (2001): The Flexible Entry. In: Zybatow, G. et al. (eds.): Current Issues in Formal Slavic Linguistics. Leipzig: University of Leipzig. 461-468.

Vitas, D, Krstev, C., Pavlović-Lažetić, G. (2000): Recent Results in Serbian Computational Lexicography. In: Bokan, N. (Ed.): Proceedings of the Symposium "*Contemporary Mathematics*", Faculty of Mathematics, University of Belgrade.

Silberztein, M.D. 2003. *INTEX Manual*, http://www.nyu.edu/pages/linguistics/intex/downloads/Manual.zip