Recent Results in Serbian Computational Lexicography

Duško Vitas[1], Cvetana Krstev[2], Gordana Pavlović-Lažetić[1], and Goran Nenadić[1]

[1] Faculty of Mathematics,
University of Belgrade
[2] Faculty of Philology,
University of Belgrade

Abstract. The recent results of the research in the construction of the electronic dictionary of Serbo-Croatian are presented. This research involves the development of methodological and theoretical principles for the construction of the lexicon for a highly inflective language. The problems that emerge on different levels of language standardization such as graphemic, orthographic, dialectic, inflective and derivational will be pointed out.

1 Introduction

Computational lexicography is a field in which results of linguistics, computer science and mathematics intersect in an object called *lexicon*. We are all familiar with the appearance of a printed dictionary, as well as with a confidence that a user feels when opening a dictionary for which he/she believes that it is a good one: the word searched for, as well as its potential meanings can be easily found. A common user can hardly observe that some information is missing from the dictionary.

The technological development does not contribute to the dictionary content: the digitized edition of a printed dictionary is usually but its transcription. Production of such dictionaries do not belong to the field of computational lexicography: editions of the Benson's Serbian-English Dictionary and Orthography dictionary of Matica Srpska on CD-ROM, both reproduce literally their corresponding traditional sources. Also, popular spelling checking programs, although they do incorporate objects that approximate words and dictionary, are not from this field either. Although undoubtedly useful, these programs profoundly rely on users' knowledge of the language.

The problems of computational lexicography begin when a text, organized by some natural language, is to be transformed without human intervention, from one computer readable form into another [29].

The inducement for the research in the computational lexicography is a search for mathematical models that can describe the stable knowledge concerning the structure of words and the grammatical system, and only

then the application of these models to the construction of natural language programming tools [6].

This article will outline the results of the research that the Natural Language Processing (NLP) group[1] at the Faculty of Mathematics has recently achieved with the aim of constructing the electronic dictionary of Serbo-Croatian[2]. This research involves the development of methodological and theoretical principles for the construction of the lexicon for a highly inflective language. Preliminary research, performed before the year 1989, that is before the concept of the integrated environment for the text processing with a dictionary as a central point in it had been introduced, will be briefly presented [36]. The problems that emerge on different levels of language standardization such as graphemic, orthographic, dialectic, inflective and derivational will be pointed out.

2 One linguistic example

We will present one example to illustrate the nature of language information that is necessary for the automatic text transformation. Many everyday users of the computer as a text entry machine have often wished to issue—instead of the common *find-replace* commands—a more elaborate command of the following form:

replace all occurrences of the noun x with the noun y

A question may be posed why such a command is never implemented. Let us suppose, as an illustration, that it is necessary to replace by only **one command** all occurrences of the noun *ugao* (Engl. *angle*) with the noun *ravan* (Engl. *plane*) in a Serbo-Croatian text. As nouns are usually represented by their nominative singular forms, such a command might have the following form:

replace all occurrences of the noun ugao with the noun ravan

Assuming that the procedure which derives all the inflective forms of nouns *ugao* and *ravan* from their nominative singular form is known, such a replacement reduces to the sequence of the following *replace* commands:

replace	*ugao*	*by*	*ravan*	(for the nominative and accusative singular forms)
replace	*ugla*	*by*	*ravni*	(for the genitive singular form)
replace	*uglu*	*by*	*ravni*	(for the dative and locative singular forms)
replace	*uglom*	*by*	*ravni*	(for the instrumental singular form)
replace	*uglovi*	*by*	*ravni*	(for the nominative plural form)
replace	*uglova*	*by*	*ravni*	(for the genitive plural form, etc.)

[1] http://www.matf.bg.ac.yu/NLP/

[2] The term Serbo-Croatian will be used in a sense introduced in [25].

The execution of such a sequence of replacement commands turns out to be useless. Namely, it does not take into account the gender of the nouns, which leads to the replacement of string *ovaj ugao* (Engl. *this angle*) with the unacceptable string **ovaj ravan* (Engl. *this plane*).[3] It is therefore necessary to assign to every string an additional information that would describe its grammatical properties.

If one would try to perform the replacement the other way round, that is to replace the noun *ravan* with the noun *ugao*, then, due to the homography of the inflective forms of the noun *ravan*, it would be necessary to analyse the grammatical information in the wider context. For instance, *u toj ravni* (Engl. *in that plane*) and *dve ravni* (Engl. *two planes*) require different forms of the noun *ugao*: *u tom uglu* (Engl. *in that angle*) and *dva ugla* (Engl. *two angles*).

The problem becomes even harder if it is necessary to replace one nominal syntagma with another one, for instance, the syntagma *pravougli trougao* (Engl. *right-angled triangle*) with the syntagma *Košijev niz* (Engl. *Cauchy's sequence*). Linguistic problems involved here are by far more complex. Both terms are compounds, i.e. nominal syntagmas with certain morphological properties: *pravougli trougao* is a syntagma which consists of the adjective *pravougli* which has only the definite form and the noun *trougao* (**pravougao trougao*), while *Košijev niz* is a syntagma in which adjective is in the indefinite form (**Košijevi niz*). Formula for string replacement would have to be enhanced in order to encompass such variations. Formula for the replacement of the genitive singular syntagmas would have the following form:

> *replace* genitive singular(*pravougli trougao*) = *pravouglog trougla*
> *with* genitive singular(*Košijev niz*) = *Košijeva niza*

Even the more peculiar behaviour can be noticed here: although for the nominative singular of the adjective *Košijev* only the indefinite form exists, in genitive this adjective occurs in definite form as well (usually it is *Košijevog niza*, and not *Košijeva niza*). However, in expressions

> *dva pravougla trougla* (Engl. *two right-angled triangles*) and
> *dva Košijeva niza* (Engl. *two Cauchy's sequences*)

both adjectives will occur only in indefinite forms.

At the same time, *Košijev niz* has the same meaning as *niz Košija* or *Košijevski niz*, so between these syntagmas the relation of synonymy can be established. Thus, the following rule can be added to the replacement command:

> *replace* possessive adjective of noun $x_{\text{sing.}}$ + noun y
> *with* noun y noun $x_{\text{sing. gen.}}$

[3] Symbol * will denote grammatically unaccaptable examples.

One could search in vain in dictionaries and grammatical manuals to find the rules that control such language processes. They are recorded neither in dictionaries nor in grammatical manuals, despite the fact that they are used in everyday speech and in written texts as well.

The presented simple example of the construction of the function *replace one noun with the other* illustrates both the goals and the tasks of the computational lexicography. What algorithms and what information make some of the stated replacements possible? The question of reliability of traditional sources cannot be avoided. We have to stress here that these problems are not singularity of Serbo-Croatian: the languages with richer linguistic tradition and more stable linguistic situation have to face them as well [3], [10], [31].

3 A short historical overview

Early works in the field of computational lexicography fall in a period before the year 1965, when one group of people worked as a test group on a Georgetown automatic translation project. This early work was interrupted by the ALPAC report (p. 42), and its negative effect was that the research continued not earlier than two decades later, first at the Mathematical institute, and then at the Faculty of Mathematics in Belgrade. This second period can be divided into two phases. During the first phase, up to the year 1990, a number of original and complex programming tools for text processing was developed, especially the system AURORA [33] and the morphological generator MORF [35]. During the second phase, the model of lexicon-grammar and the concept of electronic dictionaries were applied to Serbo-Croatian.

The main module of the AURORA system was designed so that for a given input text, marked with a logical layout tags, its internal representation was generated, which pointed to the relevant elements of the source text. Another important feature of this system was a flexible definition of a formal word: the set of separator characters was defined as input data. Based on such an internal representation, additional modules were designed that generated different text processing tools [36].

Another system produced in this period was the system MORF for morphological nominal generation, whose first version was made in 1980, and which is still in frequent use. The system MORF, for a noun or an adjective, given in a form of a dictionary entry, generates all the elements of a traditional inflective paradigm (e.g., for a noun *ugao* MORF generates the forms *ugla*, *uglu*, *ugle*, ...). This system relies upon the analysis which segments inflective words to their variable and invariable parts: for instance, the invariable part of the noun *ugao* is *ug-* and the variable part which determines the inflective behaviour of the noun is *-ao*. The noun declension affects only the variable part. For example, if variable part *-ao* corresponds to the masculine gender nouns marked as non-animated, with graphemic alternation $l \rightarrow o$ and

fleeting a, then all the nouns with these properties have the same inflective features: *ugao*, *trougao*, *vitao* (Engl. *winch*), *kotao* (Engl. *boiler*), etc. This property enables the description of the calculus over the word's variable part that performs the following transformations:

$ao \Rightarrow la$	for the genitive singular
$ao \Rightarrow lovi$	for the nominative plural, etc.

These transformations are independent of the invariable parts *ug-*, *troug-*, *vit-*, *kot-*. In order to perform this transformation, to a lexical word its **morphographemic definition** (abbreviated MGD) is assigned . The MGD contains description of the phenomena governing the process of forms generation [39]. In the case of the noun *ugao*, transformation of *ug-ao* (nominative singular) into *ug-lom* (instrumental singular) is performed through the following sequence of string substitutions over the variable part of the word:

-ao	
⇓	due to the alternation $l \rightarrow o$
-al	
⇓	due to the fleeting a
-l	
⇓	as there are no further alternations, a suffix for
-lom	the instrumental singular *-om* is concatenated

The description of the generating algorithm has not only its practical significance but it has also shown that the inflective system of Serbo-Croatian —besides its complexity—is a very stable system, in which inflective relations are precisely arranged by the morphographemic definitions.

The integration of the systems AURORA and MORF has enabled the production of a full text retrieval tool. In this system a retrieval key (lexical word) is submitted to the dictionary which generates, according to the morphographemic definition, all the forms of the key. These generated forms are matched with the dictionaries generated by the AURORA system, which then enables the extraction of all input words from a text (textual words).

Also, these experiments have led to the first attempts in the automatic concordance lemmatization [34]. In this period the first experiment was done on a query language over a textual database of biographies, with the problem of resolving ambiguities [22].

The description of the model of an integrated environment for processing linguistic data was a final result of these early experiments. The central idea of this model is that different text processing applications need to share common language resources. The performed analysis has shown that the same dictionary can be used for essentially different applications. Still, in order

to produce some real applications from these experiments, prototypes and models, a real dictionary was missing as well as a real user.

At the moment when the integrated environment for the text processing was conceived [36], the existing dictionaries of Serbo-Croatian (the dictionary [MS-MH] [18] and the uncompleted dictionary [SANU] [30]) where considered to be, due to the mentioned user's confidence, sufficient lexicographic base for the construction of the lexicon. However, the experience with the MORF system that required all the parameters to be explicitly stated in the MGD, while many of them (animatness, for instance) were missing from the traditional dictionaries, suggested that they might turn out not to be satisfactory. In 1989. the model of lexicon-grammar and the concept of electronic dictionary were adopted, both of them based on the theoretical requirements of Z. Herris [13] and developed further in LADL under guidance of professor Maurice Gross [7], [8]. The adopted models have enabled the development of critical apparatus for reconsideration of traditional lexicographic solutions and served as a methodological frame for the construction of the needed lexicon.

4 Electronic dictionaries

The system of electronic dictionaries has been described in detail in [4], [5], [9]. For this exposition of importance are the morphological dictionary of simple lexical words **DELAS** and dictionary **DELAF** which defines the relation between the textual and lexical words by the expressions of the following form:

$$W_{text},\ W_{lex}.K_n{:}K_{kat}$$

where W_{text} is an element that corresponds to a textual word, W_{lex} is an element that corresponds to a lexical word, the code K_n describes in a unique way the morphological class of inflective paradigm, and K_{kat} are the grammatical categories of the lexical word W_{lex} that are assigned to the given form of the textual word W_{text}. For instance, here is one entry in **DELAS** for French:

abandonnateur.N36

and two examples of entries in **DELAF** for French:

abandonnatrice,abandonnateur.N36:Nfs
abandonnateurs,abandonnateur.N36:Nmp

where morphological class N36 (in the system of electronic dictionaries of French) is described by sufixes *-teur*, *-trice*, *-teurs*, and *-trices*, that correspond to the marks ms, fs, mp, and fp respectively.

One of the main characteristics of the e-dictionary model is that the task of morphological analysis is replaced by the concept of lexical recognition

which means that to a textual word, by use of **DELAF**, the corresponding lexical word is attached [27], [32]. A word from a text, defined as a string of alphabetic characters delimited by separators, will be recognized as a word of the language if and only if it is found in **DELAF**. In order to implement this model, the content of traditional dictionaries that are basis of an e-dictionary construction have to be thoroughly examined [3]. The morphological class may also gather some forms of derivational paradigm which means that one lexical word, or one entry in **DELAS**, can correspond to more than one entry in a traditional dictionary.

5 Problems in the construction of DELAS/DELAF for Serbo-Croatian

Although simple in appearance, the structure of electronic dictionary of simple words cannot be directly derived from the traditional dictionaries of Serbo-Croatian for different reasons. To some of them we will point in this article. Every element in the expression W_{text}, $W_{lex}.K_n{:}K_{kat}$ is described by its own model which is not covered in the traditional grammars.

(a) The assignment of the class K_n

The class K_n that is assigned to every part of speech of simple words has to describe precisely and comprehensively all the forms of the inflective words, in particular nouns, adjectives and verbs, as well as their grammatical features (for instance, the values of grammatical categories).

In this article, we will restrict our discussion to the nominal inflection. From the computational point of view, the research in description of nominal inflection concentrates around the two opposite ideas: the **intensive** descriptions reduce the number of different paradigms to the smallest number possible, while the **exhaustive** descriptions tend to describe precisely all the possible variations. These two approaches have their own applications: the intensive description is used, mainly, in grammar manuals while in the traditional Serbo-Croatian dictionaries the exhaustive descriptions prevail. We have to stress that computer applications require strictly exhaustive description.

The formal mathematical model enables us to reduce the problem of intensive and exhaustive descriptions to the question of number of Boolean properties that are used in the definition of the equivalence relation over the set of inflective words. For instance, the intensive grammatical categorizations of noun inflection that result in two, three or four different classes use only one or two Boolean properties. The classification given in [14] represents in this respect the **minimal** intensive classification of nouns as it formally uses only one Boolean property.

In the case of exhaustive classifications [23], the number of properties is higher: for nouns, the properties covered are the noun gender, forms of genitive singular and plural, vocative singular, augmented plural forms, etc., as it is usually done in the traditional lexicography, and in consequence the set of nouns is partitioned into a higher number of classes that are not explicitly labeled. The main goal of the exhaustive class definition could be thus formalized in the following way [7, p. 214]:

"Les éléments de la classe définie par les propriétés P sont $m_1, m_2, \ldots, m_k$ (i.e. la liste des éléments), et il n'en existe pas d'autres."

For example, the inflectional paradigms of nouns *jelen*, *prozor*, *žena* are used in the system of intensive classifications as prototypes of the unmarked inflection. They can be represented by the following regular expressions:

jelen(ε/ns+*a*/gs,as,gp+*u*/ds,ls+*e*/vs,ap+*om*/is+*i*/np,vp+*ima*/dp,lp,ip) (R1)
prozor(ε/ns,as+*a*/gs,gp+*u*/ds,ls+*e*/vs,ap+*om*/is+*i*/np,vp+*ima*/dp,lp,ip) (R2)
žen(*a*/ns,gp+*e*/gs,np,ap,vp+*i*/ds,ls+*u*/as+*o*/vs+*om*/is+*ama*/dp,lp,ip) (R3)

The intensive approach implicitly uses the following definition: we say that two nouns belong to the same **inflectional class** *if the process of left factorization of their inflectional paradigms yields the same regular expression (after discarding left factors).* The resulting regular expression is denoted as **morphographemic class** (abbreviated MGC). Thus we have in the inflectional class (R1) nouns *jelen*, *aligator*, ... ; in (R2) *prozor*, *izvor*, ... ; and in class (R3) nouns *žena*, *košuta*, ... All the nouns whose regular expressions of inflectional paradigm do not correspond to the regular expressions of some intensive classification are treated as exceptions. By examining the regular expressions (R1) and (R2) we can conclude that in the intensive classification systems one regular expression is not sufficient to describe even the basic unmarked class of masculine and neutral nouns whose roots end with a consonant, as the form of the accusative singular depends on the animateness seen as a Boolean property (animate/inanimate, having as a result R1$\neq$R2).

The principle of exhaustive definition can be explained using the noun *devojka* as an example. The exhaustive description of the noun *devojka* taken from [30] is:

> **devojka** ijek. **djevojka** ž. (vok. devojko ijek. djevojko; mn. gen. devojaka (devojaka, devojki) ijek. djevojaka (djevojaka, djevojki), vok. devojke ijek. djevojke) (dijal. ek. devojća, dijal. ijek. đevojka) **1. a.** ...

For the same noun [18] gives the following description:

> **devojka**, ijek. djevojka (dijal. đevojka), ž. (dat. -ci, vok. devojko; gen. mn devojaka) **1.** ...

This noun is not in the same inflectional class as the noun *žena*, because the concatenation of the regular expression (R3) to the base *devojk-* gives the following:

$$*devojk(a/\text{ns},\textbf{?gp}+e/\text{gs,np,ap,vp}+ i/\text{ds,ls}+u/\text{as}+o/\text{vs}+om/\text{is}+ama/\text{dp,lp,ip}) \quad \text{(D1)}$$

which does not correspond to the quoted dictionary descriptions.[4] Thus, the noun *devojka* would be an exception in an intensive classification system. If one look in more details in the description given in [30], it can be seen that to **one** nominative singular form *devojka* **correspond two** different forms of genitive plural: *devojaka* and *devojki*. It would be, however, more precise to say that for **two different forms** of genitive plural, **two different forms** of nominative singular are required that have the same graphical form but to which two different sets of Boolean properties are attributed. As a result, the inflectional paradigm of noun *devojka* can be further decomposed into two regular expressions:

$$devoj(k(a/\text{ns}+e/\text{gs,np,ap,vp}+ i/\text{ds,ls}+u/\text{as}+o/\text{vs}+om/\text{is}+ama/\text{dp,lp,ip})+aka/\text{gp}) \quad \text{(D2)}$$

$$devojk(a/\text{ns}+e/\text{gs,np,ap,vp}+ i/\text{ds,ls,gp}+u/\text{as}+o/\text{vs}+om/\text{is}+ama/\text{dp,lp,ip}) \quad \text{(D3)}$$

These two regular expressions are different from (D1) in the form of genitive plural. According to [18], the factorization would be as follows:

$$devoj(k(a/\text{ns}+e/\text{gs,np,ap,vp}+u/\text{as}+ o/\text{vs}+om/\text{is}+ama/\text{dp,lp,ip})+ci/\text{ds,ls}+aka/\text{gp}) \quad \text{(D4)}$$

Every regular expression in parenthesis in (D1) to (D4) defines one equivalence class on the set of all nouns. In general, for nouns x and y the equivalence relation ρ can be defined as follows: $x\rho y$ **iff** the inflectional paradigms of x and y belong to the same inflectional class.

Every nominal inflectional paradigm belongs to exactly one such class. The same form of a nominative singular can, however, participate in more than one class: the noun *žena* belongs to the class (D1) and this class is different form the **classes** of the noun *devojka* that are described by three different expressions D2, D3 and D4.

We have to note that to the other variant forms: *djevojka*, *đevojka*, *divojka* the same regular expressions can be applied. So, they belong to the same inflectional classes as *devojka*. Applied to the description given in [30], one possible factorization could be:

[4] The question mark ? is used to indicate the regular expressions forms that are not confirmed in dictionary descriptions.

$(devoj + djevoj + đevoj + divoj)$ (D5)
$(k(a/\text{ns}+e/\text{gs,np,ap,vp}+i/\text{ds,ls}+u/\text{as}+o/\text{vs}+om/\text{is}+ama/\text{dp,lp,ip})+aka/\text{gp})$

Dialect form *devojća* does not belong to any of the classes (D2 to D4), which means that this form should have either a separate morphographemic description inside the article for *devojka* or should otherwise be a separate entry.

Relationship between the regular expression that characterizes one inflectional class and its corresponding morphographemic definition (MGD) is based on the following statements:

1. for a given regular expression of one inflectional class it is possible to construct the corresponding morphographemic definition;
2. for a given entry and its morphographemic definition, as a set of fixed values of Boolean variables, it is possible to construct the corresponding regular expression.

The first statement can be illustrated by the nouns *žena* and *devojka* and expressions (D1) to (D4) which define the values of the following Boolean variables:

a_1 = "the sibilarization in dative singular"
a_2 = "the fleeting **a** in genitive plural"
a_3 = "the suffix **-i** in the genitive plural"

The differences between the regular expressions (D1) to (D4), and their corresponding morphographemic definitions can be now represented by the table:

	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	example
D1	no	no	no	žena.N70.01
D2	no	yes	no	devojka.N70.05
D3	no	no	yes	devojka.N72.01
D4	yes	yes	no	devojka.N70.08

The second statement, applied to the noun *devojka*, has the following meaning: let the morphographemic definition with the values of parameters $\mathbf{a_1}$ = yes, $\mathbf{a_2}$ = yes, and $\mathbf{a_3}$ = no be defined for the noun *devojka*. Then, the segment *-ojka*/ns that can be automatically extracted from the string *devojka* can be computationally transformed to the strings *-ojci*/ds and *-ojaka*/gp. Factorization of the forms of the inflectional paradigm, computed in this manner, gives the regular expression D4. The similar can be applied to other combinations of Boolean properties [41].

(b) Variations of the entry $\mathbf{W}_{lex}$

The form of a lexical word which is potentially an element of the system DELAS is not stable. At least two phenomena contribute to it: dialect variations and multiple forms of the entry [16], [38].

In the morphological dictionary of SC it is necessary to neutralize the graphemic variations in the invariable part of the entries. This modification consists in use of two different alphabets: one for encoding the dictionary and another for encoding a text. One approach to this kind of redefinition of alphabet for encoding a dictionary was described in [28]. It is of particular interest to Serbo-Croatian that uses two alphabets. As the written text can be recorded using either the Cyrillic or the Latin alphabet, that do not correspond to each other unambiguously, the alphabet of e-dictionary must not depend on the alphabet used to record the source. Although the graphical variations can be historically conditioned, it is possible to effectively process them on the synchronous level using one generalization of the notion of grapheme [17].

We shall illustrate the variation of the entry using as an example the noun *hleb* (Engl. *bread*). The following regular expression represents the graphical variations in the invariable part of the entry (the part of the expression that denotes the morphological class is extracted as a common factor):

$$(h{+}\varepsilon)(l{+}lj)eb(\varepsilon/\text{ns,as}{+}a/\text{gs}{+}u/\text{ds,ls,vs} \quad \text{(H1)}$$
$$+om/\text{is}{+}e/\text{vs}{+}ov(\text{i/np,vp}{+}a/\text{gp}{+}e/\text{ap}{+}ima/\text{dp ,lp,ip}))$$

Two alternatives can be suggested here as a possible choice for an entry form. The first one is to choose for the entry the regular expression such as $(h{+}\varepsilon)(l{+}lj)eb$. Beside the introduction of a large number of regular expressions, this choice has the drawback of not providing the attributes that govern the realization of one or the other variation. As a second alternative the introduction of special graphemes called **lexicographemes**, as minimal abstract units of a dictionary encoding system was suggested. The lexicographeme must have the following property: in a text it can be realized as zero, one or more characters, where the particular choice is governed by the set of attributes assigned to the grapheme itself and by the optional operators that generalize the notion of concatenation [16]. In the example (H1), we can consider two abstract objects h', e' (thus yielding the entry form $h'le'b$): h', that can be realized as h or ε, and e', that can be realized as e or je with the property to palatalize the preceding l in some contexts. Attributes assigned to h' enable the omission of h as an unliterary or archaic possibility. Attributes are assigned to e' according to the dialect. To certain dialects the operators, such as palatalization, are assigned as well. Similarly, we can introduce lexicographeme e'', different from e', that has four different realizations: e, i, je (with the property to palatalize the preceding d), and je (with the property

not to palatalize the preceding *d*). Using e'' we can describe the entry corresponding to graphical variations of the noun *devojka* with $de''vojka$.
With lemma encoded using these new graphemes, the e-dictionary entry can be derived from the following form:

$$h'le'b(\varepsilon/\text{ns,as}+a/\text{gs}+u/\text{ds,ls,vs}+om/\text{is}+e/\text{vs} \\ +ov(i/\text{np,vp}+a/\text{gp}+e/\text{ap}+ima/\text{dp,lp,ip}))$$

The above example matches all the inflective forms of the entries *hleb*, *hljeb*, *ljeb* and *leb*.

This example expanded to the whole dictionary yields a form of an e-dictionary in which the graphical variations are treated in a strictly formal way. The resulting dictionary, although based on the dictionary obtained through the traditional process of excerption, synthesizes the graphical variations and enhances the possibilities of text retrieval. For instance, for an arbitrary entry all its graphical variations can be found (e.g. the form *hleb* returns all the inflective forms of all its graphical variations, *leb*, *ljeb*, ...). Also, the text itself need not to be altered, so its preprocessing is not necessary.

The use of lexicographemes has been limitted only to the cases where a change neither in inflectional nor in derivational class occurs. The variations produced by different replacements for "jat" are by far the most frequent. The results presented in [17, p. 169] show that in a prototype of Serbian DELAS dictionary containing 6569 entries, approximately 5% of all nouns and adjectives and 9% of all verbs were represented using 50 different lexicographemes describing the replacement for "jat".

In Serbian DELAS, lexicographemes are implemented as a reference to a canonical form (after a % sign in a dictionary entry). In order to encode the dictionary using ASCII character set only, the lexicographemes in a canonical form are represented as a reference to its character position (after a ♯ sign). This reference consists of a type of variation, character position of a lexicographeme and its type. For instance, there is an entry **beg,N07.02+*,N08.02-E%♯E2.03** in Serbian DELAS whose meaning is that string *beg* represents two nouns: *beg* (Engl. *Turkish governor*) is from the morphological class N07.02+* and has no variations, and the other is *beg* (Engl. *flight*) from the morphological class N08.02-E which has three variations described by the reference E2.03: variations are due to phonetic changes, a lexicographeme is a second character in a string *beg*, its type has a code 03. Entry for *bijeg* jek. (Engl. *flight*) refers to the same canonical form. Here are some more examples from Serbian DELAS dictionary:

bačva,N72.01-*
bežati,V31.00.3E%♯E2.12
beda,N70.01-E%♯E2.03
bedra,N70.02-*
beg,N07.02+*, N08.02-E%♯E2.03
begati,V01.00.2E%E2.12
begenisati,V21.02.4*
begluk,N04.04-*
bego,N40.01+*
begovac,N17.61+*
begunac,N17.61+E%♯E2.12
beleg,N04.02-E%♯E2.10♯E4.26
beličast,A06.01*
...
bijeda,N70.01-J%beda♯E2.03
bijeg,N08.02-J%beg♯E2.03
bijelx,N22.01-J%belx♯E2.06
bijen,A06.01*
bijes,N08.01-J%bes♯E2.03.02
bijesan,A08.52J%besan♯E2.21
...
bilxeg,N04.02-J%beleg♯E2.10♯E4.26
...
bježati,V31.00.3J%bežati♯E2.12
bjegati,V01.00.2J%begati♯E2.12
bjegunac,N17.61+J%begunac♯E2.12
bjesnxeti,V34.38.6J%besneti♯E2.20♯E5.4

(c) Form of the textual word $\mathbf{W}_{text}$

Neither the exhaustive classification of inflective classes nor the neutralization of variations on the level of meta-dictionary are enough to cover all the possible forms of textual words. These problems emerge from multiple sources. On one side, for some entries morphological system allows for more than one inflective class to be chosen. However, neither traditional excerption applied in the production of dictionaries nor the grammatical manuals give reliable information about the inflective properties of the entries. This problem was discussed in [40]. Our approach to the solution of this problem is based on a view, somewhat different from usual, to the *hapax legomenon*. A criteria is adopted that a class can be assigned to an inflective word only in conjunction with a corpus which can confirm a critical form in its paradigm. For instance, for a feminine gender noun ending with a consonant, in order to assign to it a class code, it is sometimes necessary to know, beside the form of instrumental singular, also the form of genitive plural. Without confirmations of these critical forms in coprus, it is not possible to assign reliably to a lexical word its class code and thus they cannot be included in system DELAS. One different approach has been applied in [1]: instead of posing the above constraint, the free variation of parameter values that determine the critical forms is allowed. For instance, for the feminine gender nouns ending with a consonant both suffixes for that the morphological system allows, $-i$ and $-ju$ (with the note where palatalization applies), are equally cited. This approach permits the free systematic variations of the parameters in the MGD, without any constraints that would possibly be imposed by corpus. The similar approach has been discussed in [37]. Beside the incosistency in implementation of this approach, it has the further negative consequence of degrading the concept of MGC.

The second important problem is the phenomenon of structural derivation which, when applied to an entry, does not produce the change of the lexical meaning of the entry. In system DELAS the relation existing between

the noun *jelen* and, for instance, its diminutive forms *jelenče* or *jelenčić*, and derived adjectives *jelenov* or *jelenji* is not represented. Beside that, the number of entries in DELAS is significantly increased. The other problems that arise from this organization of morphological dictionary are analyzed in [2], [11], [24]. We suggest one solution to this problem that is based on the extension of the regular expression of the form (R1) that encompasses the forms, which are the result of the structural derivation. For instance, in the following expression

$$jelen(\varepsilon/\mathrm{N1}+če/\mathrm{N2.Dim}+čić/\mathrm{N3.Dim}+ov/\mathrm{Adj1}+ji/\mathrm{Adj2}+ski/\mathrm{Adj3}) \quad \text{(R5)}$$

every member in the parenthesis represents the appropriate regular expression having the form similar to (R1). The regular expression of the form (R5) can be derived either by analyzing the derivation processes on corpus (as is recorded in [21]), or by projecting the derivational system to the dictionary content (as was done in [1], for instance). The extension of regular expression has several important consequences:

1. The entry in DELAS dictionary becomes the canonical representative of the family of words having the same lexical meaning, without the loss of information on the type of the derivational process. This significantly reduces the number of lexical units that have to be processed in constructing and maintaining the electronic dictionary.
2. Due to the structure of the regular expression (R5) it is possible to reproduce also the entry formats that are used in traditional lexicographic practice enabling at the same time more effective formal processing of the corpus.
3. The information contained in the regular expression enables, on the level of compound words, the treatment of certain types of noun phrases as synonymous despite the different nature of their grammatical structure [11]. For instance, the strings *kći konzula* (Engl. daughter of consul) and *konzulova kći* (Engl. consul's daughter) can be reduced to a common canonical form using the transformations on the level of regular expressions (for instance, by transforming the adjective into the noun in genitive).

Finally, a textual word need not necessarily correspond to the class that is assigned to the appropriate lexical word [26]. Sometimes class is not obtained by means of MGD—namely, it can be influenced by some morphosyntactic parameters which are not included in MGD.

6 An application of the electronic dictionary

The suitability of the described basis for the construction of electronic dictionary of simple words in Serbo-Croatian has been tested in the project aimed

at the production of the electronic edition of the collection of proverbs compiled by Vuk S. Karadžić [15]. The inventory of proverbs in all paper editions and the forms they are presented in have not changed essentially since the first edition in 1849. The last edition thus maintains the old orthography of the original as well as the elements of the Old Church Slavonic alphabet, for instance hard sign, and stressed letters.

The new paper and electronic edition should have had to be accompanied by an inteligible index of key words for text retrieval. On the basis of text concordances the index of key words in text has been compiled as well as the meta-dictionary encoded using lexicographemes. From this meta-dictionary, the dictionary DELAS was generated first and then the dictionary DELAF, and they enabled the automatic lemmatization of concordances which was independent of the possible variations in text. This form of DELAS dictionary together with the standard form of DELAF dictionary can be used to unify the variations. This can be illustrated with one excerpt from the new concordances where the keyword is the canonical form of lemma:

delo♯E2.39.N
<pv>Jače je **djelo** nego besjeda.</pv>

deo♯E3.04.10.N
<pv>Jadan je onaj koji nema **dijela** od Boga.</pv>

doneti♯E4.02.V
<pv>Ja poslah sina u Rim da primijeni turin, a on kad dođe iz Rima, **donese** dva turina.</pv>

drem♯E3.03.N
<pv>Ja kad viđeh zelen drijen, predadoh mu vas moj **drijem** i lijen.</pv>

dren♯E3.07.01.N
<pv>Ja kad viđeh zelen **drijen**, predadoh mu vas moj drijem i lijen.</pv>

goreti♯E4.13.V
<pv>Jedna palica ni pred carem ne **gori**.</pv>

greh♯E3.03.04.N
<pv>Jedna šteta sto **grijeha**.</pv>

hadumac♯H1.07.N
<pv>Ja mu kažem **adumac** sam, a on pita koliko đece imam!</pv>

7 Current activities and further improvments

After the successful experiment with the electronic edition of Vuk's Proverbs, the described principles have been applied to the construction of the system of morphological dictionaries of simple words. At present, dictionary DELAS has approximately 70.000 entries in which the lexicographemes have been only

partly encoded (particularly those that reflect the old 'yat'). The process of revision of entries is performed through the interaction with the corpus.

The initial experiments with the construction of the system of dictionaries of compound words (DELAC/DELACF) have also been successfully accomplished [19]. First attempts in the construction of local grammars and their testing with the system INTEX [32] have been undertaken [20].

Acknowledgment

We would like to express our gratitude to Prof. Maurice Gross and his group from LADL, for all the support and information that have been of significant help in our research.

References

1. Anić, V.; 1998: *Rječnik hrvatskoga jezika*, (treće prošireno izdanje), Novi Liber, Zagreb
2. Clemanceau, D.; 1993: *Structuration du lexique et reconnaissance de mots dérivés*, Thèse de doctorat d'informatique fondamentale, Université Paris 7, LADL, CERIL, Avril 1993
3. Courtois, B.; 1989: *Construction du lexique DELAS: Codificaition et contrôle des entrées lexicales*, LADL, Paris, mai 1989
4. Courtois, B.; Silberztein, M. (eds.); 1990: *Dictionnaires électroniques du français*, Langue française, 87, Larousse, Paris, septembre 1990
5. Courtois, B.; 1990: *Un système de dictionnaires électroniques pour les mots simples du français*, in [4], pp. 11-22
6. Gross, M.; 1972: *Mathematical Models in Linguistics* , Englewood Cliffs N.J.: Prentice Hall Inc.
7. Gross, M.; 1975: *Méthodes en syntaxe*, Hermann, Paris
8. Gross, M.; 1988: *Methods and Tactics in the Construction of a Lexicon-grammar*, The Linguistic Society of Korea (ed.), Linguistic in the Morning Calm 2, Seoul; Hanshin Publishing Co. pp. 177 - 197
9. Gross, M.; 1989a: *The Use of Finite Automata in the Lexical Representation of Natural Language*, in: Gross, M.; Perrin, D. (eds.): *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science, no. 377, Springer-Verlag, Berlin, pp. 34–50
10. Gross, M.; 1989b: *La construction de dictionnaires électroniques*, Annales des télécommunications, 44(1–2), pp. 4–19
11. Gross, M.; 1997a: *Synonymie, morphologie dérivationnelle et transformations*, Langages, no. 128, Larousse, Paris, pp. 72-90
12. Gross, M.; 1997b: *The Construction of Local Grammars*, Roche, E.; Schabes, Y. (eds.): Finite State Language Processing, Cambridge, Mass.: The MIT Press, pp. 329-352.
13. Harris, Z.; 1964: *The Elementary Transformations*, Philadelphia: University of Pennsylvania, TDAP N 54. Reprinted in Papers in Structural and Transformational Linguistics, 1970, Dordrecht: Reidel, pp.482-532.

14. Ivić, P.; 1990: *Sistem padežnih nastavaka imenica u srpskohrvatskom književnom jeziku*, in: *O jeziku nekadašnjem i sadašnjem*, BIGZ - Jedinstvo, Beograd - Priština
15. Karadžić, Vuk, S.; 1997: *Srpske narodne poslovice sa indeksom ključnih reči*, Nolit, Beograd
16. Krstev, C.; Vitas, D; Pavlović-Lažetić, G.; 1995: *Neutralization of Variations in the Structure of a Dictionary Entry in Serbo-Croatian*, Formale Slavistik, Junghanns, U.; Zybatow, G. (eds.); Vervuert Verlag, Frankfurt am Main, pp. 417–425
17. Krstev, C.; 1997: *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije*. Doktorska disertacija, Matematički fakultet, Beograd
18. MS-MH, 1968: *Rečnik srpskohrvatskoga književnog jezika*, vol. 1–6, Matica srpska, Matica hrvatska, Beograd-Zagreb
19. Nenadić, G.; 1997: *Algoritmi prepoznavanja složenih reči u matematičkom tekstu i njihove primene*, magistarska teza, Matematički fakultet, Beograd, p. 151
20. Nenadić, G.; Vitas, D.; 1998: *Using Local Grammars for Agreement Modeling in Highly Inflective Languages*, in Proc. of First Workshop on Text, Speech, Dialogue - TSD 98, Brno, pp. 91-96
21. Nenadić, G.; Spasić, I.; 1999: *The Acquisation of Some Lexical Constraints from Corpora* , in Text, Speech and Dialogue - TSD 99, Lecture Notes in Artificial Intelligance 1692, Springer-Verlag, pp. 115-120
22. Pavlović-Lažetić, G.; 1987: *Baze podataka i ekspertni sistemi u upravljanju tekstom*, Doktorska disertacija, Matematički fakultet, Beograd
23. Piper, P.; 1994: *Alternacijski tipovi i kongruencijske klase u srpskoj imeničkoj paradigmi*, Zbornik Matice srpske za filologiju i lingvistiku, XXXVII/1-2, Novi Sad, pp. 499-510
24. Popović, Lj.; 1996: *Morphosyntactic strings*, Studije srpske i slovenske, serija I, godina 1, Srpski jezik broj 1-2, Naučno društvo za negovanje i proučavanje srpskog jezika, Beograd, pp. 90-101
25. Popović, Lj.; 1996: *Deux approches idéologiques de la vernacularisation de la langue littéraire chez les Serbs à la fin du 18e et dans la premiére moitié du 19e siécle*. Langues et nation en Europe Centrale et Orientale du 19e siécle à nos jours, Cahiers de l'ILSL, no. 8: 209-240. Lausanne.
26. Popović, Lj.; 1999: *Bivalentni kontrolori kongruencije (kao problem gramatičkog i leksikografskog opisa)*, saopštenje na Naučnom sastanku slavista u Vukove dane, MSC, Beograd
27. Roche, E., Schabes, Y. (ed.); 1997: *Finite-State Language Processing*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England.
28. Sampson, G.; 1989: *How Fully Does a Machine-Usable Dictionary Cover English Text?*, Literary and Linguistic Computing, vol. 4, No. 1, pp. 29-35
29. Salton, G.; 1989: *Automatic Text Processing (The Transformation, Analysis, and Retrieval of Information by Computer)*, Addison-Wesley Publ. Comp, Reading, Massachusetts
30. SANU: *Rečnik srpskohrvatskog književnog i narodnog jezika*, vol. 1–14 (A–N), Srpska akademija nauka i umetnosti, Institut za srpskohrvatski jezik, Beograd, 1959–1990
31. Silberztein, M.; 1989: *Dictionnaires électroniques et reconnaissaince lexicale automatique*, Thèse de doctorat en Informatique fondamentale, Université Paris 7, novembre 1989

32. Silberztein, M.D.; 1993: *Le dictionnaire électronique et analyse automatique de textes: Le systeme INTEX*, Masson, Paris
33. Vitas, D.; 1979: *Prikaz jednog sistema za automatsku obradu teksta*, Zbornik radova XIV jugoslovenskog međunarodnog simpozijuma o obradi podataka Informatica 79, Bled, Slovensko društvo INFORMATIKA, Ljubljana, p. 7 101
34. Vitas, D.; 1979: *Jedan postupak automatske segmentacije srpskohrvatskih reči i njegove primene*, Zbornik III konferencije "Computer Processing of Language Data", Bleda, 1985, pp. 303-313
35. Vitas, D.; 1981: *Generisanje imeničkih oblika u srpskohrvatskvom jeziku*, Informatica 3/81, Ljubljana, pp. 49–55
36. Vitas, D.; Tancig, P.; 1988: *Skice za izgradnju integrisanog ambijenta za obradu tekstuelnih informacija* , Zbornik IV konferencije "Computer Processing of Language Data", Institut "Jožef Stefan", Portorož, pp. 1-12
37. Vitas, D.; Krstev, C.; 1992: *Interaction between Dictionary and Text in Serbo-Croatian*, in: Ferenc Kiefer, Gabor Kiss, Julia Pajzs (eds): Papers in Computational Lexicography, COMPLEX'92, Budapest, pp. 333-342
38. Vitas, D; Pavlović-Lažetić, G.; Krstev, C.; 1993: *Electronic Dictionary and Text Processing in Serbo–Croatian*, in: Józef Darski; Zygmunt Vetulani (eds.): Sprache — Kommunikation — Informatik, Max Niemeyer Verlag, Tübingen, pp 225–231
39. Vitas, D.; 1993: *Mathematical Model of Serbo-Croatian Morphology (Nominal Inflection)*, PhD thesis, Faculty of Mathematics, University of Belgrade
40. Vitas, D.; Krstev, C.; 1996: *Tuning the Text with an Electronic Dictionary*, in: Ferenc Kiefer, Gabor Kiss, Julia Pajzs (eds): Papers in Computational Lexicography, COMPLEX'96, Budapest, pp. 267–276
41. Vitas, D.; 1997: *O elementarnoj morfografemskoj klasi*. Naučni sastanak slavista u Vukove dane, no. 26/2, MSC, Beograd, pp. 195-206