1. Introduction

In recent years, morphological dictionaries for Intex (M. Silberztein, 1993) were developed for some Slavonic languages (K. Bogacki, 1997), (S. Koeva, 2000), (L. Meskova, 2000), (D. Vitas, 2000). As they were constructed independently and based on national conceptions of grammatical description of morphophonology and morphology, as well as on results of national lexicography, their methodological bases, scopes and grammatical tags differ significantly. However, they have a certain number of common features at the morphological level, reflecting similarity of their inflective and derivational systems.

In the frame of comparative studies of Slavonic languages we find diachronic studies with no direct relevance for construction of morphological electronic dictionaries. Scarce contrastive studies provide frames for methodologically uniform description of parts of morphological systems of Slavonic languages on synchronic principles (T.V. Popova, 1987), (H. Dalewska-Gren, 1997). But even in such studies, as to the best of our knowledge, we cannot find elements that would enable standardization of descriptions according to the requirements of an electronic dictionary construction.

Such a situation has at least two consequences: a) analysis of different - however close - Slavonic languages by the Intex system does not necessarily give comparable results and b) introduction of a new Slavonic language into the Intex family is confronted with problems met and solved in the process of construction of other electronic dictionaries.

The purpose of this paper is to suggest possible methodological solutions for the construction of morphological dictionaries of Slavonic languages, based on the experience acquired in the process of developing DELAS/DELAF dictionaries for Serbian language.

2. Similarities of morphological systems

Common morphological features of Slavonic languages have their historical explanations in proto-Slavonic and Old Church Slavonic sources, as is, for example, the case with development of the system of morphonological alternations. Diachronic relations enable establishment of useful analogies.

As an example of morphological similarity, we will present the realization of the noun *writer* in different Slavonic languages. It is a noun of masculine gender; its nominative form is given below (1).
The example illustrates the nature of structural similarity of different Slavonic languages (though our intention is not to suggest the general conclusion that such segmentation is always possible or even desirable).

Declension of above-mentioned nouns in respective morphological systems is characterized by a system of endings. We will present the simplified examples of declension tables, reduced to the form of regular expressions, for some of the nouns from the above example:

(2) a.  Russian: \textit{pisatel}' (/mns+ja/mgs,mas+ju/mds+je/mls+jem/mis +i/mnp+jej/mgp,map+jam/mdp+jami/mlp+jah/mlp)

b.  Polish: \textit{pisarz}(<E>/mns+a/mgs,mas+owi/mds+u/mvs/mls, +em/mis +e/mnp,mvp+ow/mgp,map+om/mdp+ami/mlp+ach/mip)

c.  Slovak: \textit{spisovatel}'(/mns+a/mgs,mas+ovi/mds,mls+om/mis,mdp +ia/mnp+ov/mgp,map+och/mls+mi/mip)

d.  Slovenian: \textit{pisatelj}(<E>/ mns+a/mgs,mas+u/mds,mls+em/mis +i/mnp,mip+ev/mgp+ema/mdp+e/map+ih/mlp)

e.  Serbian/Croatian: \textit{spisatelj}(<E>/mns+a/mgs,mas,mgp+u/mds,mls,mvs+em/mis+ +i/mnp,mvp+e/map+ima/mdp,mlp,mip)

The tags for gender (m for masculine, f for feminine, and n for neuter), case (n - nominative, g - genitive, etc.) and number (s -singular, p - plural) are listed behind the slash.

Similarities of morphological systems illustrated by (1) and (2) suggest that in the construction of a DELAF dictionary it is possible to apply similar methods for these languages. First of all, in the above regular expressions, the right factor can be presented
as FST. For instance, the right factor for \textit{spisatelj} in Serbo-Croatian can also be described by a graph designated N17.01.

\[
\begin{array}{c}
\text{<spisatelj> N17.01} \\
\end{array}
\]

On the other hand, in the above regular expressions, the right factor has classificatory role. Namely, it is possible to define the relation of "inflective" equality in the following way:

Two variable entries \( x \) and \( y \) belong to the same inflective class if and only if the right factors in the realization of their inflective paradigms are equal.

For example, according to this definition of "inflective" equality, Serbian nouns \textit{uciteli}, \textit{spasitelj}, \textit{kucxepazitelj}, \textit{graditelj} are inflectively equal with the noun \textit{spisatelj} because they have the same right factor described by the graph N17.01, i.e.
The right factor in a regular expression unambiguously describes inflective behaviour of inflected words and enables an exhaustive and precise description of inflective classes, which is a prerequisite for the construction of a DELAF-type dictionary. The right factor is not a result of a random factorization. As is demonstrated in (Vitas, 1997), some individual members of the right factor can be generated from the entry form on the basis of morphographemic definition which, expressed as a set of Boolean variables, provides an explanation of relations inside an inflective paradigm. In other words, the right factor has its "explanation" that corresponds to the way in which traditional grammar interprets inflective processes.

3. Construction of a DELAF dictionary

Proceeding from regular expressions obtained by factorization of inflective paradigms (1), a DELAF-type dictionary can be constructed in many ways.

The basic way is to construct graphs on the basis of regular expressions, and to apply the procedure described in (M. Silberztein, 2000) as in the example of the application of the graph N17.01 to the nouns from the class to which spisatelj belongs. To each inflective class then corresponds exactly one graph.

Intex enables some other approaches. Let us consider the following example. In some classes of nouns in Serbian, an infix is inserted between stem and endings, while the set of endings remains the same as in a class without such an infix. Here are some examples of such models in Serbian:

(3)  
   a. Infix -ev- in plural:  
      ns: puzx (Engl. snail)  
      gs: puzx+<E>+a;  
      np: puzx+ev+i
   b. Infix -in- in singular:  
      ns: gra|anin(Engl. citizen)  
      gs:gra|an+in+a;  
      np: gra|an+<E>+i

Such cases can be processed by replacing the graph N17.01 with two graphs that separately describe endings for singular and plural: singN17 and pluralN17 as presented in the figure:
Nouns without singular or plural form, as well as nouns in which the change of number leads to the change of gender can be treated in similar way:

   b. ns: *dinamo (Engl. dynamo)- masc np: *dinama - neuter or *diname - fem.

In cases (3) and (4), a precisely established set of endings is added to the fixed stem, so that each such graph defines one inflective class. A problem arises when graphemic alternations entail a change of stem. The noun *pisac (1.e) in Serbian:

(5) \( pi(sac(<E>/mns+a/mgp)+sc(a/mgs,mas+...+i/mnp+...)+\{sxcye/mvs\}) \)

has the same meaning and set of endings as the noun *spisatelj, but is inflectively different from it because the factorization of its inflective paradigm gives a different right factor. Inflective difference is produced by the occurrence of a fleeting 'a in the nominative singular. Hence, its genitive singular form is *pisca instead of *pisaca, which means that to generate the forms from the expression (5) the graph N17.01 cannot be directly applied.

Modified base can be dealt with in several ways. One way is to use stack-operator implemented into Intex. The other is not to use the entry itself, but its left factor (produced by the factorization of inflective paradigm), as is the segment 'pi in (5). Such cases then appear as subclasses of the base class. The noun *picas will then pertain to the class N17.12, differing from the class N17.01 only in certain set of graphic alternations (fleeting 'a and palatalization in the vocative singular). This approach is applied for generation of inflective paradigms in Serbian, by using function \texttt{re2fst} (in Intex/App) that is not integrated into INTEX-graphical environment. This function for a given regular expression of the inflective class produces the FST, which applied to the list of entries belonging to that class, produces an output which, with certain lesser modifications, presents an uncompressed dictionary in DELAF-format. For example, the inflective class of the verb *pevati (Engl. to sing) is described by the following regular expression:
Application of the FST produced by the function `re2fst` to the verb described in DELAS by `pevati.V1+Imperf+Tr+Iref+Ek` results in the following:

(7)  pevati/:W,pevati.V1+Imperf+Tr+Iref+Ek
    pevam/:Pxs,pevati.V1+Imperf+Tr+Iref+Ek
    pevasx/:Pys,pevati.V1+Imperf+Tr+Iref+Ek
    peva/:Pzs:Ays:Azs,pevati.V1+Imperf+Tr+Iref+Ek
                      .........................
    pevasxe/:Azp:Iys:Izs,pevati.V1+Imperf+Tr+Iref+Ek
                      .........................

This procedure proved to be especially suitable when the regular expression defining a class is too complex. For example, passive (past) participle of the verb `pevati` - `pevan` (Engl. sung) - besides being used for passive voice may also function as an adjective, and then it has adjectival inflective paradigm. Regular expression (8) describes a part of this paradigm where endings of the past participle are set in bold, while the rest (after `/`) are the associated values of grammatical categories:

(8)  <E>/:akmsNNv:akmsNNq:akmsAAq +
    a/:akmsGGv:akmsGGq:akmsAAv:akmsNNv:akfsNNv:akfsNNq:akfsNNq:akfsVVv:akfsVVq:akfsVVv:akfsVVq:akfsVVq:akfsVVq:
    q:aknpVVq +
                      .............................................................................................................................
    q:akfpVVq:
                      .............................................................................................................................

(Doubled letters (NN, GG, etc.) are in final version of DELAF replaced by digits 1, 2, etc. that represent the corresponding case).
This means that the simple word form `pevati - pevana` can represent 25 different grammatical forms.
4. Some problems

If a precise and complete description of inflection has been accomplished, and a good dictionary is available, the above procedure makes possible almost automatic construction of a DELAF-type dictionary. However, if mentioned conditions are not fulfilled, we are confronted with problems demanding a careful reconsideration of results of traditional morphological and lexicographical descriptions. Let us point out some of the problems arisen in the process of construction of a morphological dictionary of Serbian.

4.1. Problems at the inflectional level belong to two basic groups: those originating in possible variations in the left and in the right factor of a factorized inflective paradigm, that is in the entry and in its inflective class.

4.1.1. Variations in the entry. An inflective word may have several forms of entry. Nouns and adjectives may have multiple forms of the nominative singular (e.g. as a result of dialect variations in Serbian, the noun *devojka* - Engl. *girl* - has three forms of the nominative singular: *devojka, đevojka, defojka*). Similarly, a verb may have multiple forms of the infinitive: (e.g.: *voljeti, voljeti* - Engl. *to love*). In traditional lexicography they are recorded as separate entries. However, all dialect variants of an entry belong to the same inflective class. In this case, we are faced with possibility of choice. We can regard each variant as a separate entry in DELAS. It may be effective, but yields two problems: (a) the number of entries in DELAS grows beyond control, (b) the existing semantic relation between these forms is broken. The other possible way, outlined in (D. Vitas, C. Krstev, 1999) uses a special way of encoding for DELAS.

4.1.2. Multiple classes. Unlike the previous case, where for one entry several dialect variants exist which all belong to the same inflective class, there are cases where more inflective classes correspond to one entry. This phenomenon occurs when an entry has more alternative endings for a single grammatical form (e.g. the noun *devojka* has two forms for the dative singular, as well as two forms for the genitive plural). Two cases are possible here: (a) alternative forms are equal in value and it is necessary to form a union of corresponding regular expressions; (b) alternative endings may indicate a change of meaning. Then it is necessary to split an entry into two or more entries, each having its own regular expression.

It is also possible that separate variants of an entry have different inflective classes:

(9) a. *litar*, m (N02.01) = *litra*, f. (N72.01) (Engl. *litre*)
    b. *definisati, defini*em (V21) = *definirati, definiram* (V1) (Engl. *to define*)

In such cases, a property suggesting possible alternative form of the entry is introduced in DELAS of Serbian. Verbs from (9b) are registered in DELAS as:

(10) *definisati, V21+Perf+Imperf+Iref+DerSatiRati*
    *definirati, V1+Perf+Imperf+Iref+DerRatiSati*
where parameters $\text{DerSatiRati}$ and $\text{DerRatiSati}$ indicate an alternative form of the infinitive.

4.1.3. Hierarchy of forms. Unlike the French dictionary, where hierarchy was established within DELAS (levels 1, 2, 3) with a view to reduce ambiguity, the Serbian language requires hierarchization of certain forms within the right factor. For example, certain forms of imperfect or aorist (with law frequencies in contemporary texts), as well as the vocative of nouns, often produce homography. Possibility to exclude such forms from the indexing process significantly reduces ambiguity.

4.2. Problems at derivation level. Another group of problems arises from the fundamental question how to restrict an inflective paradigm. For Serbian these limits are usually fixed by strict distinction between derivation and inflexion. Thus, for example, pairs marked by difference in sex/gender (e.g. $\text{spisatelj}$, m versus $\text{spisateljica}$, f (Engl. woman writer); or $\text{lav}$, m versus $\text{lavica}$, f (Engl. lioness)) are treated as separate (although closely related) entries. This problem is outlined in (D. Vitas, 2002) for cases where a change of POS does not occur and a solution is suggested that extends the relation of inflective inequality (gender motion, amplification of meaning).

The problem of relation between derivation and inflexion is deeper; it encompasses those derivational phenomena in which the meaning of an entry obtained by derivational process is closely related to the meaning of the base word (or predictable from it), but in which a change of part of speech occurs.

Comparing results of alignment of the original text of Voltaire's *Candide ou l'optimisme* with its Serbian translation, we can observe that the entry $<$baron$>$ (Engl. baron) in the French text has 56 occurrences compared to 45 in the Serbian translation. The difference is due to the fact that the French class N/A41, to which $<$baron$>$ belongs, covers more inflective classes in Serbian, i.e.
Hence, if the Serbian text is retrieved by the pattern "super-baron", corresponding to the union:

(11) \(<\text{baron}>+<\text{baronica}>+<\text{baronov}>+<\text{baroncyin}>\)

The "super-entry" has the same number of occurrences as in the French original, which is illustrated by the following examples from concordances of the aligned text:

(12)
(1.1.1) en Westphalie, dans le château de M. le baron de Thunder-ten-tronckh, un jeune
(1.1.1) u Vestfaliji, u zamu gospodina barona Tunder-ten-tronka,
(1.3.1) Madame la baronne, qui pesait environ trois cent cinquante livres, s'attirait par là
(1.3.1) Gospodxa baronica bila je tesxka, otprilike, tri stotine pedeset funti i time je pribavljala
(1.3.3) Le fils du baron paraissait en tout digne de son père.
(1.3.3) Sin baronov izgledao je u svemu dostojan svog roditelja.

Likewise, the analysis of aligned text of Verne's Journey around the World in 80 Days gives:

(13) \(N \rightarrow \text{AdjPoss}\)
(4.23.3) Il prit le sac des mains de Passepartout.
(4.23.3) On uzme torbu iz ruku Paspartuovih.
(6.20.2) ... minarets se dessinaient au-dessus de la ville sous les pâles rayons du soleil.
(6.20.2) ... minareta ocrtavalo se iznad grada pod bledim suncyevim zracima.

\(V \rightarrow \text{Gerund (Verb noun)}\)
(1.10.3) Son seul passe-temps était de lire les journaux et de jouer au whist.
(1.10.3) Jedina mu je zabava bila cyitanje (N) novina i igranje (N) vista.

These examples demonstrates that, when part of speech is changed due to this kind of derivational processes, the basic meaning of words remains the same; they also show that, in majority of examples, base and derived word may be substituted without disrupting the sentence structure (M. Gross, 1997).

The question arises how to deal with these kind of derivational phenomena in the frame of DELAF? One possible way is suggested in (D. Vitas, C. Krstev, 1999): instead of regular expressions formed as in (2), extended regular expressions could be formed with a view to gather these different entries into a "super-entry" (as in the example of <baron>). This treatment has both its advantages and disadvantages: on one hand, comparability of different dictionaries and possibility to gather close forms into a single entry in DELAS are achieved, while, on the other hand the "unusual" way of lemmatization is obtained with significant increase in the number of classes because derivational suffixes are combined without mutual dependencies.

The other solution for the problem of a form generation of a "super-entry" may be realized by a method suggested in (M. Gross, 1989) and illustrated by the well-known graph France.grf. Such procedure is applied to the construction of a Serbian verb subsystem DELAF: the verb in Serbian may have the form of passive participle with a twofold function. Its nominative forms are part of the conjugation of (mainly) transitive
verbs and they participate in the construction of passive voice composite tenses (6). But, the passive participle behaves independently like other adjectives, which means that it has all the usual case forms (8). Likewise, imperfective verbs are used in the construction of a gerund having a nominal inflective class.

Verbs are generated according to the following procedure. First, for a group of verbs with the same conjugational properties the corresponding elements of DELAF are generated. Then the nominative singular form of masculine gender of the passive participle and the base of gerund are extracted by a separate function; in the next step, to thus generated nouns and adjectives the corresponding FSTs are applied that generate their inflective paradigms.

Indicators are added, using syntactic markers in DELAS, which indicate that adjective and nominal paradigms generated in this way represent the passive participle, i.e. the gerund.

\[ \text{podavijati}, V1+\text{Imperf}+\text{Tr}+\text{Iref} \]
\[ \text{podavijan}, A1+\text{PP} \]
\[ \text{podavijanxe}, N300+\text{VN} \]

Similar approach may be applied to the outlined problem of entries derived from a noun. Such procedure procures information about the nature of derivational process by which a given derived entry was produced; still, the information about its ancestor-entry is lacking.

References

Bogacki Christophe; 1997. POLLEX - un dictionnaire électronique morphologique du polonais, Actes du colloque international FRACTAL'1997, BULAG, Besançon

Courtois, Blandine; Max Silberztein (eds.).1990. Dictionnaires électroniques du français.
Langue française 87. Paris: Larousse

Gross, Maurice; 1989. The Use of Finite Automata in the Lexical Representation of Natural Language; in: Gross, Maurice; Dominique Perrin (Eds.). Electronic Dictionaries and Automata in Computational Linguistics, Lecture Notes in Computer Science. Berlin: Springer Verlag


Koeva, Svetla; Mihov, Stoyan; 2000. Intex 4.0 for Bulgarian, 3es Journées INTEX, Liège


Popova, T.V. (ed.); 1987. Slavjanskaja morfonologija (Substantivnne slovoizmenenie), Moskva: Nauka


Silberztein, Max D. 2000. INTEX, Paris: Asstril


Vitas, Duško; Krstev, Cvetana; 2000. Recent Results in Serbian Computational Lexicography. In: Bokan, Neda (Ed.): Proceedings of the Symposium "Contemporary Mathematics", Faculty of Mathematics, University of Belgrade

Dusko Vitas
Faculty of Mathematics
Studentski trg 16
YU-11000 Beograd

Cvetana Krstev
Faculty of Philology
Studentski trg 3
YU-11000 Beograd