Cvetana Krstev, Duško Vitas & Gordana Pavlović-Lažetić, Belgrade

Resources and methods in the morphosyntactic processing of Serbo-Croatian

1. Introduction

The precise morphosyntactic processing of Serbo-Croatian, having in mind its complex morphology, depends on the availability of annotated electronic lexical resources, in the sense of ISO 1087-2 (2000). At the same time, results of morphosyntactic processing always depend on how precise and detailed the lexical resources are. This paper describes some of the lexical resources for the morphosyntactic processing of Serbo-Croatian that have been developed in the scope of the network RELEX, as described in Laporte (2003). They constitute mainly the system of a morphological electronic dictionary and local grammars implemented using the technique of finite state transducers in the INTEX programming environment, described in Silberztein (1993). The resources developed in this way allow various language experiments and the refinement of results in text analysis.

In section 2 of this paper we will present the lexical resources for the processing of Serbo-Croatian, while in section 3 we will present the methods for morphological and morphosyntactic processing based on these resources and finite state transducers. Various applications using the presented resources and methods are illustrated with several examples.

2. Lexical resources

Following RELEX recommendations, resources for the morphological processing of Serbo-Croatian comprise:

- Morphological dictionaries of simple words DELAS/DELAF
- Morphological dictionaries of compounds DELAC/DELACF
- Morphological grammars - dictionary of lexical transducers DLM
- Local grammars

2.1 Dictionaries of simple words

Simple words are defined as alphabetic strings, or character strings between delimiters, where delimiters are all the characters that do not belong to the alphabet of a language as explained in Vitas (2001). A simple words module consists of three parts:

- Dictionary of lemmas DELAS
- Set of transducers describing the properties of inflectional paradigms
- Dictionary of inflected forms DELAF

All the entries in the DELAS dictionary have the following form:

\[ \text{lemma}.\text{Knnn} [+\text{SinSem}]^* \]

where \( \text{lemma} \) is the simple word, in general in the form usually used in traditional dictionaries, \( K \) is the part of speech mark, \( nnn \) is the number denoting the class of lemmas that all share the same inflectional properties described by the appropriate transducer \( \text{Knnn} \), and \( +\text{SinSem} \) is the freely attached marker that describes the syntactic, semantic, derivational, or other properties of a lemma (Table 1).

A part of speech code, e.g. N, and an inflectional class code, e.g. 688, uniquely determine the finite transducer that generates all the forms in a lemma paradigm. A finite transducer, being capable of producing the output, adds to all these forms their possible grammatical categories. A regular expression describing the inflectional paradigm N688 is
given in Figure 2. This regular expression applied to the noun of the class 688 (for instance, propoved ‘sermon’) generates three different forms of the noun (propoved, propovedi, propovedima) each of which has several possible grammatical realizations. The first of them propoved, for instance, represents the inanimate (q) feminine gender (f) form which is in the singular (s), in nominative (1) or accusative case (4) (Table 2).

<table>
<thead>
<tr>
<th>Category</th>
<th>tag</th>
<th>applied to</th>
<th>explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>+p2</td>
<td>preposition</td>
<td>noun phrases in genitive</td>
<td>bez,PREP+p2</td>
</tr>
<tr>
<td></td>
<td>+Ref</td>
<td>verbs</td>
<td>reflexive</td>
<td>dicyiti,V551+Imperf+It+Ref</td>
</tr>
<tr>
<td></td>
<td>+MG</td>
<td>nouns</td>
<td>masculine natural gender</td>
<td>budala,N601+Hum+MG+FG</td>
</tr>
<tr>
<td>derivational</td>
<td>+VN</td>
<td>nouns</td>
<td>verbal noun</td>
<td>kiselnxenxe,N300+VN</td>
</tr>
<tr>
<td></td>
<td>+Adj</td>
<td>adverbs</td>
<td>derived from adjectives</td>
<td>fanaticyno,ADV+Adj</td>
</tr>
<tr>
<td></td>
<td>+DerOvaIra</td>
<td>verbs, nouns, adjectives</td>
<td>derivational variety</td>
<td>dezinfikovati,V18+...+DerOvaIra</td>
</tr>
<tr>
<td>Semantic</td>
<td>+Col</td>
<td>adjectives</td>
<td>colors</td>
<td>zelenkastosiv,A6+Col</td>
</tr>
<tr>
<td></td>
<td>+Hum</td>
<td>nouns</td>
<td>human</td>
<td>lxubavnica,N601+Hum</td>
</tr>
<tr>
<td></td>
<td>+Mat</td>
<td>adjectives</td>
<td>material</td>
<td>kozxnat,A6+Mat</td>
</tr>
<tr>
<td>Dialectic</td>
<td>+Ek</td>
<td>all</td>
<td>croatism</td>
<td>izopcxen,A1+PP+Cr</td>
</tr>
</tbody>
</table>

Table 1. Some markers attached to lemmas in Serbian DELAS dictionary

Figure 1. An excerpt from Serbian DELAS dictionary

```
<Ex>/:fs1q:fs4q +  
i/:fs2q:fs3q:fs5q:fs6q:fs7q:fp1q:fp2q:fp4q:fp5q +  
ima/:fp3q:fp6q:fp7q
```

Figure 2. The regular expression N688 that describes the inflectional properties of the noun class to which belongs peć ‘stove’, cev ‘pipe’, noć ‘night’, etc.

The DELAS dictionary and the set of transducers describing inflectional properties are used to produce the DELAF dictionary of all word forms. All the entries in the DELAF dictionary have the following form:

```
form,lemma[:categories]*
```

where `form` is a simple word form of a `lemma` that is represented by its DELAS entry form, and `:categories` are the possible grammatical categories of the word `form`, each category
represented by the single character code explained in Table 2. The set of grammatical categories that describe the word form depends on the part of speech, and, in some cases, on the form itself. For instance, for verbs, the impersonal forms infinitive, present gerund and past gerund are represented by just one code, \( W \), \( S \), and \( X \) respectively, active past participle and passive past participle by codes \( G \) and \( T \) respectively together with codes representing the gender and number, while other simple tenses are represented by the tense code, number code, and person code. As a certain number of Serbian nouns have a different gender in a different number, this is reflected in the set of grammatical codes attached to the inflected forms. For instance, two forms with associated grammatical codes of the noun \( \text{sudija} \) ‘judge’ are \( \text{sudija:ms1v:ms5v:fp2v} \) and \( \text{sudije:ms2v:fp1v:fp4v:fp5v} \). An excerpt from the Serbian DELAF dictionary is given in Figure 3.

<table>
<thead>
<tr>
<th>PoS</th>
<th>category</th>
<th>codes</th>
<th>PoS</th>
<th>category</th>
<th>codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun – N</td>
<td>gender</td>
<td>m - masculine</td>
<td>adjective - A</td>
<td>gender</td>
<td>case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>f - feminine</td>
<td></td>
<td></td>
<td>number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n - neutre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>case</td>
<td>1 - nominative</td>
<td></td>
<td></td>
<td>animatness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 - genitive</td>
<td></td>
<td></td>
<td>degree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 - dative</td>
<td></td>
<td></td>
<td>a - positive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 - accusative</td>
<td></td>
<td></td>
<td>b - comparative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 - vocative</td>
<td></td>
<td></td>
<td>c - superlative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 - instrumental</td>
<td></td>
<td></td>
<td>d - definite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 - locative</td>
<td></td>
<td></td>
<td>k - indefinite</td>
</tr>
<tr>
<td></td>
<td>number</td>
<td>s - singular</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>p - plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>animatness</td>
<td>v - animate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>q - non-animate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>g - no consequence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>verb - V</td>
<td></td>
<td></td>
<td></td>
<td>W - infinitive</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P - present</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A - aorist</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I - imperfect</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y - imperative</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G - active PP</td>
</tr>
<tr>
<td></td>
<td>preposition - PREP</td>
<td>/</td>
<td></td>
<td></td>
<td>T - passive PP</td>
</tr>
<tr>
<td></td>
<td>conjunction - CONJ</td>
<td>/</td>
<td></td>
<td></td>
<td>F - future</td>
</tr>
<tr>
<td></td>
<td>particle - PAR</td>
<td>/</td>
<td></td>
<td></td>
<td>S - present gerund</td>
</tr>
<tr>
<td></td>
<td>Interjection - INT</td>
<td>/</td>
<td></td>
<td></td>
<td>X - past gerund</td>
</tr>
<tr>
<td></td>
<td>Numerals - NUM</td>
<td>case</td>
<td>person</td>
<td>x - first</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>number</td>
<td></td>
<td>y - second</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gender</td>
<td></td>
<td>z - third</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>animatness</td>
<td></td>
<td>h - negated clitic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pronouns - PRO</td>
<td>case</td>
<td>negation</td>
<td>i - positive clitic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>number</td>
<td></td>
<td>number</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>animatness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>clitic</td>
<td></td>
<td>i - clitic</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Partial list of morphological categories and their codes used in Serbian electronic dictionaries.
As semantic markers do not form a closed set, they can and are modified and expended in the course of the dictionary development. For instance, some verbs are now marked with +Ref as potentially reflexive verbs. However, the distinction between true reflexive verbs such as smejati se ‘to laugh’, non-true reflexive verbs as kupati se ‘to bath oneself’ and reciprocal verbs as voleti se ‘to love each other’ has not yet been done.

Some of the markers are exclusive, meaning that depending on realization one or the other can stand, while others are not. For instance, presently one verb can be marked both as transitive +Tr and intransitive +It, e.g. brisati,V21+Imperf+Tr+It+Iref; the transitivity of such verbs depends on the usage. In the sentence Rekord koji je postigla sxtafeta brisan je sa liste svetskih rekorda (‘The record accomplished by the relay team has been deleted from the list of world records’) verb brisati ‘to delete’ is transitive, while in the sentence To telo dahcx,e, cvili, pisxti kao stara parnxacya, brixe, savija, puzi (‘The body gasps, whines, whistles like an old steam locomotive, blows, folds, creeps’) it is intransitive.1 This suggests that such verbs in a future and for particular applications can be treated as two lemmas. An excerpt from the Serbian DELAS dictionary illustrating the usage of +SynSem markers is given in Figure 1.

At present, the Serbian DELAS dictionary consists of approximately 70,000 entries from which nearly one million entries in DELAF are generated. These dictionaries can be applied to an arbitrary text using the Intex programming system, which then enables retrieval from the text of occurrences, using any criteria that can be expressed in terms of dictionary content, such as particular lemmas, particular parts of speech, the realization of specific grammatical categories, word forms with specific syntactic or semantic features, as well as their combinations. The following examples will illustrate some of these possibilities.

**Example 1.**

Dictionaries and grammars state that nouns belonging to the large class of primarily abstract feminine nouns N704, for instance aktivnost ‘activity’ and nezavisnost ‘independence’, have two singular forms in the instrumental case: aktivnosxcxu and aktivnosti, nezavisnosxcxu and nezavisnosti. The following query

```plaintext
<PREP+p6> <A:fs6>* <N+N704:fs6>
```

retrieves from the text all the single word form sequences that start with the preposition taking the instrumental case, followed by the optional sequence of adjectives in the singular, feminine gender, and instrumental case, and ending with a feminine noun

---

1 All the examples in this article have been extracted from the corpus of contemporary newspaper and literary texts and processed with Intex.
belonging to the class N704 in the singular instrumental case\(^2\). An excerpt from the produced concordances is given in Figure 4. First form of the instrumental case occurred in a number of cases (Figure 4a). All cases that superficially look like the realization of the second form of the instrumental case turn out to be the realization of some other case, genitive or accusative, since many prepositions can be used with more than one case.

Figure 4. Part of the concordance lines produced by the query <PREP+p6> <A:fs6>* <N:fs6>. a) The use of nouns from the class N704 in instrumental case. b) Prepositions that can be followed by the instrumental case, with noun forms that can be in instrumental singular. However, other cases are realized.

Example 2
Agreements with numerals (see Leko (2003)) can be analyzed on corpus. The complex query

\[(<\text{NUM}> <\text{N+Hum:m}> (<V:G> + <V:T>) \text{jesam}) + (<\text{jesam}> (<V:G> + <V:T>) <\text{NUM}> <\text{N+Hum:m}>)\]

tries to retrieve from a text certain constructions that illustrate the agreement of the quantitative nouns (number \(<\text{NUM}>\) followed by a noun \(<\text{N}>\) of masculine gender denoting humans \(<\text{N+Hum:m}>\) with an auxiliary verb \(<\text{jesam}>\) and active or passive past participle \(<\text{V:G}>\) or \(<\text{V:T}>\) respectively). The produced concordances show that nouns quantified by numbers equal to or greater than 5 agree with the auxiliary verb in the singular and past participle in paucal, coinciding in form with the neuter singular, while nouns quantified with numbers less than 5 agree with the auxiliary verb in plural and the past participle in feminine singular (Figure 5a). However, as the query itself does not take into account the agreement, among the concordance lines there are a number of false retrievals (Figure 5b). Means to reduce the production of non-relevant concordance lines will be introduced in section 3.1. Also, the query

\[<\text{N+NumN}> <\text{N+Hum:m}> <\text{ADV}>* \text{jesam} (<\text{V:T}> + <\text{V:G}>) + <\text{N+NumN}> <\text{N+Hum:m}> (<\text{V:T}> + <\text{V:G}>) \text{jesam} <\text{ADV}>*\]

retrieves from the text some usages of numeral nouns \(<\text{N+NumN}>\) that show their agreement with auxiliary verbs and past participles. The retrieved occurrences show that these nouns, although used only with the nouns of masculine gender, sometimes agree with the past participle in feminine (Figure 5c).

---

2 Special markers, such as +N704 are included only in the version of the dictionary for researchers; other versions contain no references to the inflectional classes.
Figure 5. Some concordance lines illustrating the agreement of: a) the agreement of quantified nouns; b) overproduced concordance lines; c) numeral nouns.

2.2 Special dictionaries

Beside the DELAS and DELAF dictionaries, which contain the general language lexica, various special dictionaries are also being constructed. Two the most developed at this moment are the dictionary of toponyms and other geographic names DELAS-TOP and the dictionary of Serbian personal names DELAS-PERS as described in Pavlović-Lažetić (2004), Grass (2002). The entries, both lemmas and forms, have the same format as in general dictionaries — the only difference is in the usage of a different set of semantic markers. The most prominent marker in a DELAS-TOP dictionary are +Top for toponyms, +Hyd for hydronyms, and +Oro for oronyms (see Figure 6a), while in DELAS-PERS +First is used for first names, +Last for surnames, and +Nick for nicknames (see Figure 6b).

Figure 6. DELAS entries in special dictionaries. a) Toponyms: Beograd is a proper name (+NProp), toponym (+Top), a capital city (+PGgr) in Yugoslavia (+IsoYU). b) Personal names: Varnicyicx is a proper name (+NProp) for a human (+Hum), surname (+Last).

Some of the syntactical markers assigned to the entries in DELAS special dictionaries are related in the WordNet-like manner, as discussed in Krstev (2004), that is one marker can signify a concept that is a hypernym of a concept signified by the other marker. For instance, Sava is a proper name (+NProp), more precisely a hydronym (+Hyd), and still more precisely a river (+PRek). When retrieving information from a text, different levels of abstraction can be used, depending on the task at hand.

Example 3.
The query \(<N+First> <N+Last> + <N+Last> <N+First>\) iz <N+Top> retrieves from a text all complete personal names (both first name and surname in an arbitrary order) denoting a person from ('iz') a certain place (<N+Top>). The result of such a query is given in Figure 7. This query is, however, rather naive, since it does not take into consideration the agreement between first name and surname (see section 3.1).

Figure 7. Retrieved concordance lines with the query \(<N+First> <N+Last> + <N+Last> <N+First>\) iz <N+Top>. This experiment shows that the order first name followed by surname prevails since no other order occurs once in a text of 1MW.

2.3 Dictionary of compounds
Compounds, as defined in Silberztein (1993), are similar to simple words except that they can contain separators, that is non-alphabetic characters, as well. What makes them different from simple words is that each simple word making a compound can inflect itself, following its own inflectional rules. On the other hand, compounds are similar to simple words in that respect, since no inserts are allowed between the simple word constituents. Also, the syntactic and semantic markers can be associated to the compound, as it represents a distinct concept (Figure 8).

The recognition of compounds by using the appropriate dictionary is important since it can add additional information to the morphologically analyzed text, that can be useful in the syntactic analysis. For instance, the sequence na vreme is a prepositional phrase in the sentence Priprema se muzicyki program u znak podsecxanxa na vreme kada je poznati pesnik ucyitelxevao u tom selu (‘A musical programme is under preparation in remembrance of the time when the famous poet was a teacher in that village’). However, the same sequence is very often the compound adverb, for example in the sentence Dali smo gol na vreme, ali na kraju nismo istorxili nekoliko sxans (‘We scored a goal on time but in the end we missed several chances’). The difference between two usages is that simple word constituents in a compound adverb never change, while the noun vreme ‘time’ can be used also in plural in the prepositional phrase, for instance, Da li me je vecx taj pasus podsecxao na vremena koja su prosxla? ‘Did that paragraph itself already remind me of times that had passed?’

Figure 8. A few entries from the DELACS dictionary of compounds

Although the limit between, for instance, noun compounds and free nominal groups is not always easy to establish, in many cases it is clear that the noun compound: (a) has its own rules of inflection; (b) does not allow inserts; and (c) represents a concept that differs from the concepts of its constituents. The compound Ujedinxene nacije ‘The United Nations’ illustrates the point (a). Namely, only if used in plural and with constituents in the given order can it represent the institution (compare to za nacije ujedinxene ovim sporazumom ‘...for the nations united by this treaty’). Also, when another adjective is inserted in the sequence odgovorno lice between the adjective odgovoran ‘responsible’ and the noun lice ‘person’ the noun compound odgovorno lice ‘responsible person, e.g. head, boss’ is not realized in the text. For instance, Vlada je iz budjeta uplatila 11,5 miliona dinara jednom broju novinskih kucxa i odgovornih fizicykih lica (‘The Government has paid 11.5 million diners to a certain number of newspaper houses and responsible physical persons’). Finally, the literal meaning of lice is ‘face’, thus it represents a part of the body, which means that its semantic marker in DELAS could be +BodyPart. However, preceding it with an adjective various compounds are obtained: odgovorno lice ‘responsible person (boss, head)’, svesxteno lice ‘clerical person (clergyman, reverend)’, pravno lice ‘juridical person (firm, house)’, vojno lice ‘military person (military man, serviceman)’, all of which would obtain their own semantic markers: for the given examples all could obtain the marker +Hum for human being, except pravno lice, which could obtain the marker +Org for organization. More prominent is the example of the compound pokondirena tikva conceited pumpkin (upstart, parvenu), where simple and compound nouns would be semantically marked as:

tikva,N+Bot
pokondirena tikva,N+C+Hum
The dictionary of inflected forms of compounds DELACF can be constructed automatically from DELACS in a similar way to the dictionary of inflected simple forms DELAF being constructed from the dictionary of simple lemmas DELAS (Figure 9). However, two types of transducers are needed for this task: the first are transducers used to produce all the inflected forms of simple words (see Figure 2) and the second are transducers that govern the agreement between the constituents in a particular compound. This second transducers are not implemented in Intex itself, so the production of Serbian DELACF is for the time being produced outside the system. In the future, the formalism for the description of these second transducers, as well as the corresponding format for entries in DELACS, will be adopted as it is developed by Savary (2000).

![Figure 9. An excerpt from the Serbian dictionary of inflected compound forms DELACF](image)

**Example 4**

The usage of personal pronouns and particularly their clitic forms, based on the examples from Milićev (2003), can be analyzed on corpus as well. The query `<PREP+p2> <PRO+Prs:2>` retrieves all personal (+Prs) pronouns (PRO) in the genitive case (2) that are preceded by a preposition (PREP) which requires the genitive case (+p2). Both simple (bez 'without', blizu 'close to', do 'next to', kod 'at') and compound prepositions (za razliku od 'contrary to', u okviru 'in the scope of', as described in Popović (1966)) are taken into consideration. The produced concordance entries confirm that the full pronouns and not the clitics follow the preposition (Figure 10a). The query `<PREP+p2> <PRO+Prs:2i>` produces no relevant concordance lines (i denotes the clitic forms). Similarly, the query `<> <PRO+Prs:2>` retrieves all personal pronouns in the initial position (<>) in the linguistic unit (usually, a sentence), showing again that only full pronouns can be found in this position (Figure 10b). As before, the query `<> <PRO+Prs:2i>` does not yield relevant results. These and similar results can be used to construct disambiguation tools (see section 3.3).

![Figure 10. Some concordance lines produced by the query](image)

3 Finite-state automata and transducers

3.1 Finite state automata and regular expressions
The examples from section 2 have illustrated how regular expressions can be used to retrieve certain occurrences from texts. However, regular expressions can easily become very complex (see Example 2) and difficult to understand, correct, and augment. In such cases it is more appropriate to use the finite state automata (FSA), for which in Intex an editor is provided, which provides a convenient way to formulate a query. Moreover, the developed finite state automata can be named and used by that name in other automata.

**Example 5**

Retrieval precision in Example 2 can be improved if finite state automata are used instead of regular expressions, in which the agreement constraints can be more easily expressed. Figure 11 represents a graph that is an enhancement of the first regular expression from Ex. 2: besides taking care of agreement, it also allows for the insertion of an adverb between the auxiliary verb and past participle. Some produced concordance lines are given in Figure 14a.

![Figure 11. A simple model of the grammatical agreement between quantitative masculine gender nouns and corresponding predicates.](image)

In Stanojčić (2000, sect. 492) it is stated that besides grammatical agreement, the semantic agreement of quantitative nouns with the predicate is also possible; that is the predicate agrees both in number and gender with the quantitative noun. A graph of FSA that retrieves from a text agreement of this sort is given in Figure 12. However, in a sample corpus of newspaper and agency texts of more than one million words, the examples of such agreement were not found. Nonetheless, if this FSA is augmented, as shown in Figure 13, to encompass relative clauses as well, examples of the semantic agreement can be extracted (Figure 14b). If the FSA from the Figure 11 is augmented in a similar way to cover relative clauses, the examples of the grammatical agreement of quantitative nouns and predicate in a relative clause can be also detected (Figure 14c).
3.2 Finite state transducers for text modification and transformation

A local grammar is a finite transducer that on the basis of lexical resources, enables the extraction of complex structures from a text. The extracted structures can be defined using some formal criterion (for example, ‘identify all the occurrences in the text referring to an inhabitant of Belgrade visiting Greece during the year 2002’), or according to morphosyntactic, syntactic or semantic criteria. One example of such a local grammar is the automaton that recognizes, tags and lemmatizes composite tenses in Serbian, described in Vitas (2003).

Example 6.
A cascade of FSTs Vreme is produced that recognizes in a Serbian a denotation of time periods and points in time and encloses them in the XML start and end tags <period> and <datum> respectively. The nodes in the main FST in this cascade (Figure 15) represent the use of other FSTs, each of which recognizes certain class of patterns. One of these subgraphs is given in Figure 16. Such graphs can be used for text preprocessing, namely for the automatic XML tagging (Figure 17).
Roughly speaking, the finite state transducers (FST) are finite state automata that can produce the output. In text processing, this output can be used to transform the original text either by replacing the recognized pattern by the defined output or by inserting it in the specified position in text.

Figure 16. The subgraph trenutak called by the FST from Figure 15. It recognizes the points in time representing formal denotation of date (DanDatum) as well as some more informal denotations.

Figure 17. A few concordance lines of a text tagged with the FSTs Vreme.

3.3 Finite state transducers for disambiguation

One simple or compound word form is ambiguous if and only if it has been associated to several different lexical information after applying one or more lexical resources. The level of ambiguity in Serbian texts is very high since there are a lot of homographs that are not homophones. Roughly speaking, more than two thirds of running words in Serbian text are ambiguous if all the grammatical information is taken into account, and more than one fifth of them have more than one lemma potentially associated to them. Visually, the ambiguity of a text can be represented by the text graph (Figure 18) in which separate nodes exist for each running word to which several lemmas were associated in the course of lexical recognition.
The ambiguous text blurs the results of various text analysis tasks performed using such a text and makes the further phases of processing, as syntactic analysis, difficult, if not impossible. Ambiguity can be automatically resolved using different methods, the aim of all of them being not to introduce errors in a tagged text:

- **Filter dictionaries.** The applied dictionaries can be ranked, some of them given a higher priority than others. The options suggested by these dictionaries have precedence over the options found in other dictionaries, and thus resolve the ambiguity. For instance, if an entry *li,.PAR* is put in the highest priority dictionary, then it will override the possibility *li,liti.V:Ays:Azs* suggested by some other dictionary, that is, the word form *li* will always be treated as a particle *li*, and the possibility that it is some aorist form of the verb *liti* ‘to pour’ will be rejected.

- **Dictionary of compounds.** The dictionaries of compounds usually add to the ambiguity of a text, since many compounds are used in the literal meaning as well (see the example *na vreme* from the section 2.3). However, if used with the higher priority they can act as filters themselves. For instance, the entry from the dictionary of compounds of high priority *leptir-masxnom,leptir-masxna.N:fs6q* enforces the following interpretation: if the noun *leptir* ‘butterfly’ in nominative singular is followed by a hyphen which is followed by a word *masxna* ‘bow’ in any inflected form than the whole sequence should be interpreted as a compound *leptir-masxna* ‘bow tie’ in the same inflected form.

- **Disambiguation graphs.** Local grammars can be constructed with the aim of eliminating the ambiguity of certain words by specifying some of the context in which they appear. These grammars can be formulated in the form of transducers that recognize these particular contexts in order to destroy the overflow of lexical hypotheses. These are achieved by associating the lexical constraints to the identified sequences.

**Example 7**

From the sentence graph in Figure 18 it can be seen that the word forms *ni* and *do*, among others, are ambiguous: the first can be both particle (‘no’) and conjunction (‘neither/nor’), the second a preposition (‘to’) and an adverb (‘only; approximately’). However, in constructions such as this one, where the preposition (*do* in this case) is used with the negative pronoun (*nisxta* ‘nothing’), the word forms *ni* and *do*, can be unambiguously interpreted. This interpretation is enforced by the FST with constraints given in Figure 19. The application of this FST on text yields the modified sentence graph with some of the ambiguities resolved (Figure 20).
3.4 Transducers with lexical constraints

The special transducers with lexical constraints, the so called lexical transducers, allow the expression of morphological rules that govern word formation. The input of lexical transducers is used to recognize word forms, while the output is used to compute the corresponding lemma and other grammatical information. They can be quite complex and can perform the tokenization of word forms into linguistic units. These linguistic units are established on the basis of imposed constraints which are expressed in terms of recognition by e-dictionaries. Furthermore, during the recognition process the values of the recognized linguistic units can be stored in variables, which can later be used for the computation of lemmas and grammatical categories.

Example 8

Many entries in Serbian traditional dictionaries are due to a structural derivation that yields the predictive meaning of a lemma. In a dictionary such entries are recognized by
their prototype definitions. Moreover, they are usually derived using some recognizable derivational pattern. Numerous lexical transducers have been produced that make use of these derivational patterns to recognize possessive adjectives (Figure 21), diminutives, augmentatives, gender motion, prefixation, etc, analyzed in Pavlović-Lažetić (2004). To the word forms recognized by these lexical transducers the same information is associated as the word forms in the DELAF dictionary itself (Figure 22).

**Figure 21.** The lexical transducer that recognizes the possessive adjectives reduced by suffixes -ov and -in from masculine gender nouns are in a DELAS dictionary of simple nouns. The lexical graph takes into account the morphological transformations fleeting -a (arbitrov, a possessive adjective of arbitar ‘arbiter’) and the substitution of a voiced consonant with its unvoiced counterpart (pretkov, a possessive adjective of predak ‘ancestor’). A dummy entry w.A1+Dummy is used to recognize the inflectional endings of a possessive adjective.

- a) cyovekovim, cyovekova.N+Pos:aems6g:aemp3g:aemp6g:aemp7g:aefp3g:aefp6g:aefp7g...
dxindxickeve,dxindxickeve:aemp4g:aefp4g:aefp5g
    Firerov,Firerov:akms1g:aems4q
    javorovog,javorov:adms2g:adms4v:adms2g
- b) adolescentkinxa,adolescentkinxa.N+Hum:fs1v:fp2v
    bolnicyarke,bolnicyarka.N+Hum:fs2v:fp1v:fp4v:fp5v
    bolesnice,bolesnica.N+Hum:fs2q:fs5q:fp1q:fp4q:fp5q
- c) antiratnih,antirati.A2+PosQ:aemp2g:aefp2g:aenp2g
cyetrnaestogodisxakcyetrnaestogodisxak,N10+Hum:mp1v:mp5v
dvopiopstruki,dvopiopstruki.A6:adms1g:adms4q:aems4g:aemp1g:aemp5g
- d) dvadesetcyetvoro,dvadesetcyetvoro.NUM06+Coll:ns
dvadesetcyetvoro,dvadesetcyetvoro:fp1v:fp4v:fp5v
    kofercyicxa,kofercyicx.N+Dem:fs1v:fp2v
- e) cxercyica,cxercyica.N+Dem:fs1v:fp2v
    kofercyicxa,kofercyicx.N+Dem:ms2q:imp2q

**Figure 22.** The word forms recognized by lexical transducers: a) possessive adjectives; b) feminine gender counterparts of masculine nouns; c) prefixed adjectives and nouns; d) arbitrary collective numbers; e) diminutives.
4 Conclusion

Quality of morphosyntactic processing of Serbo-Croatian written texts highly depends on electronic lexical resources available as well as on methods for morphological and morphosyntactic processing. The electronic lexical resources developed and presented in the paper consist of the system of morphological electronic dictionaries of simple words, compounds and specialized dictionaries of geographic and personal names, as well as local grammars implemented using finite state transducers technique in the programming environment INTEX. Processing methods presented include finite state techniques for retrieving occurrences on complex queries, for text modification and transformation, disambiguation, as well as lexical transducers for expressing morphological rules that govern word formation such as possessive adjectives, prefixed adjectives and nouns, arbitrary collective numbers, diminutives, etc.

Although not yet completed, lexical resources described in the paper provide for fulfillment of a broad class of morphosyntactic processing tasks, illustrated by examples.

References


Savary, A. (2000). Recensement et description des mots composés - méthodes et applications. Thèse de doctorat en Informatique Fondamentale, Université de Marne-la-Vallée


