A Description of Morphological Features of Serbian: a Revision using Feature System Declaration

Cvetana Krstev[1], Ranka Stanković[2], Duško Vitas[3]
[1] professor, Faculty of Philology, Belgrade, [2] assistant professor, Faculty of Mining and Geology, Belgrade
[3] professor, Faculty of Mathematics, Belgrade
E-mail: cvetana@matf.bg.ac.rs, ranka@rgf.bg.ac.rs, vitas@matf.bg.ac.rs

Abstract

In this paper we discuss some well-known morphological descriptions used in various projects and applications (most notably MULTEXT-East and Unitex) and illustrate the encountered problems on Serbian. We have spotted four groups of problems: the lack of a value for an existing category, the lack of a category, the interdependence of values and categories lacking some description, and the lack of a support for some types of categories. At the same time, various descriptions often describe exactly the same morphological property using different approaches. We propose a new morphological description for Serbian following the feature structure representation defined by the ISO standard. In this description we try do incorporate all characteristics of Serbian that need to be specified for various applications. We have developed several XSLT scripts that transform our description into descriptions needed for various applications. We have developed the first version of this new description, but we treat it as an ongoing project because for some properties we have not yet found the satisfactory solution.

1. Motivation

Description of morphological features of a language is a prerequisite for many NLP applications. This description can be simple or complex depending both on a language and application in question. Considerable efforts in standardizing such a description were undertaken, but many applications ignore them, sometimes because they do not meet their needs. For Serbian, the most important standardized description of morphological features MULTEXT-East (Erjavec, 2004) was used in several projects (Kešelj et al., 2004), (Popović, 2009). Serbian morphological dictionaries of simple words and compounds developed in the LADL format (Courtois & Silberztein, 1990) use different morphological description (Krstev, 2009), but it was shown that both formats are to certain extent interchangeable (Krstev et al., 2004). However, some authors have already pointed to deficiencies of MULTEXT-East description for Slavic languages (Przepiórkowski, 2003). On the other hand, several applications developed in the frame of LADL e-dictionary format use their own morphological descriptions, most notably ELAG for morphological disambiguation (Laporte & Monceaux, 1999) and Multiflex for compound inflection (Savary, 2008). When our own system for automatic detection of compound inflectional properties (Krstev & Vitas, 2009) asked for yet another description of morphological features, we realized the necessity of producing a comprehensive description that could be easily transformed to any format needed by particular applications.

2. Inadequacies of Existing Descriptions

We will illustrate inadequacies of existing morphological descriptions for Serbian on MULTEXT-East, serving as *de facto* standard for involved languages for many years. They can be grouped in several categories.

(a) The lack of a value for an existing category. This problem occurred, for instance, with the grammatical category number. The values for this category for nouns are singular and plural (shared by all involved languages), and besides that dual (for Slovene and Czech) and count (for Bulgarian). This last value in Bulgarian is similar to 'paukal' in Serbian which is used with small numbers two, three and four. This value, however, does not exist for the category number for verbs although it should, as adjectives and nouns agree in number for certain verb forms. For instance,

> *... dva čoveka(paukal) su se šetala(paukal) na keju usred gomile domorodaca i stranaca koji nagrću u taj grad…*
>
> *Two men were promenading up and down the wharves, among the crowd of natives and strangers who were sojourning at this once straggling village…*

This problem was in the past solved by adding the missing values in the new MULTEXT-East versions.

(b) The lack of a category. In Serbian, gender of nouns is a grammatical category, and it has values: masculine, feminine and neuter. The value of this category is different in certain cases from the natural gender (or sex) which affects some agreement conditions. For instance, in the next example *petorica* 'five men' has the grammatical gender feminine and natural gender masculine, and in the same time grammatical number singular and natural number plural. Natural gender and number, instead of grammatical gender and number are used for verbal forms *sastali* and *su*.

> *To veče, petorica(fem/sing) džentlmenovih kolega sastali(masc/pl) su(pl) se već u osam časova u velikom salonu Reform-kluba.*
>
> *The five antagonists of Phileas Fogg had met in the great saloon of the club.*

The solution to this kind of problem asks for addition of new categories that was not encouraged by MULTEXT-East in the past.

(c) The interdependence of values and categories. In

Serbian, value paukal for the category number differs from other values. Namely, values singular and plural exist for all nominal cases, while this is not the case for 'paukal' which has a fixed form attributed to the genitive and accusative cases. Also, most of the cardinal numerals does not inflect *pet*, *šest*, *sedam* … 'five, six, seven…', some may inflect in case (but need not) *dva*, *tri*, *četiri* 'two, three, four', some inflect in gender *dva* 'two', and some inflect obligatory in gender, number and case *jedan* 'one'. This information is in MULTEXT-East represented by lists of possible combinations of categories and their values. Such representation is perhaps not the easiest to maintain and re-use, especially when interdependences exist between two or more categories. For instance, in Serbian the form of an adjective in masculine singular, in accusative case depends on the animacy of the corresponding noun; for all other combinations of values for gender, number and case animacy is of no importance.

(d) The types of categories. MULTEXT-East does not make an explicit difference between category types. Some attributes are fixed for a chosen lemma, like for instance, a type of a noun that can be either common or proper. The other categories inflect freely for all parts-of-speech (PoS) for which they apply, like the grammatical case for nouns, adjectives, pronouns and numerals. Some categories inflect for some PoS and remain fixed for other: the grammatical gender does not inflect for nouns but inflects for adjectives. However, all categories are not binary: inflects vs. fixed. For instance, certain Serbian nouns change gender in plural, like *izdajica* 'traitor' and *knjigovođa* 'bookkeeper': they have grammatical gender masculine in singular, but grammatical gender feminine in plural. Such information need not be important for some applications (like tagging) but can be for the other (generation).

3. One Illustrative Example

To stress the points stated above, we will illustrate how the problem of 'paukal' is solved in morphosyntactic descriptions used by various applications. It differs from other values of the category number since it does not exist for all inflectional cases. This is not the unique property of nouns and 'paukal'—the formally similar situation in which only a subset of values and categories can combine often arises in many languages. For instance, in Serbian, forms for the imperative mood for verbs exist only for the second person in singular and the first and second person in plural.

(a) In MULTEXT-East v. 3.0 (and previous versions) necessary information can be given in the form of a combination table. Although such tables were not supplied for Serbian, the combination table describing the 'paukal' for nouns could have the following form:

Type	Gen	Nb	Case	Anim
any	any	[sp]	any	any
any	any	t	g	any
any	any	t	a	Any

(b) Multiflex, as a part of the Unitex (Paumier 2008), that we use for the inflection of Serbian compounds and the production of e-dictionary of compound forms uses a very simple morphosyntactic description. The following line describes the inflection of Serbian nouns:

```
noun:(Nb,<var>),(Case,<var>),
     (Gen,<fixed>),(Anim,<fixed>)
```

It states that categories number and case, (previously introduced in the description) inflect for nouns and can take all previously defined values, while categories gender and animacy are fixed. For the purpose, this description is sufficient because the inflectional process is governed by inflectional transducers.

(c) The Elag subsystem of Unitex for resolving ambiguities uses much more elaborated morphosyntactic description. Its description of 'paukal' is similar to MULTEXT-East:

```
    <gender> s <case> <anim>
    <gender> p <case> <anim>
    <gender> w 2 <anim>
    <gender> w 4 <anim>
    Const <gender> s <anim>
```

All categories used for nouns and their values are previously defined in the description. (Value w used for 'paukal' corresponds to the value t in MULTEXT-East, while values 2 and 4 correspond to g and a, respectively.) The last line in the above description gives important additional information: nouns marked by a semantic marker +Const in an e-dictionary do not inflect (e.g. feminine personal names of foreign origin *Karmen* and numeral nouns *dvadesetak* 'approximately twenty').

(d) As both Elag and Multiflex are used as Unitex subsystems, a new morphosyntactic description in the form of a XML document was produced.[1] In this new format, all information presented in sections (a) to (c) concerning nouns (including 'paukal' and nouns that do not inflect) is represented as follows:

```
<inflectional-features>
   <feat att="Nb" variable="true"/>
   <feat att="Gen" variable="false"/>
   <feat att="Case" variable="true"/>
   <feat att="Anim" variable="false"/>
</inflectional-features>
   <syntactic-semantic-features>
      <feat att="sem1" val="Const"/>
   </syntactic-semantic-features>
   <feature-combinations>
      <comb><feat att="Nb" val="s"/>
             <feat att="Gen"/>
             <feat att="Case"/>
             <feat att="Anim"/></comb>
             …
      <comb><feat att="Nb" val="w"/>
             <feat att="Gen"/>
             <feat att="Case" val="2"/>
             <feat att="Anim"/></comb>
         …
      <comb><feat att="sem1"/>
             <feat att="Nb" val="s"/>
```

[1] This descriprion was produced by Eric Laporte and Agata Savary (personal communication December, 2009)

```
        <feat att="Gen"/>
        <feat att="Anim"/></comb>
    </feature-combinations>
```

4. The New Solution

Our main goal was to produce a new systematic morphological description of Serbian: (1) that would not deviate much from the traditional description; (2) that would be compatible with already developed morphological dictionaries; (3) that would be compatible with MULTEXT-East description; (4) and that would be reusable for present and future applications. Thus we have developed SprFSD following the feature structure representation defined by the ISO 24610 standard (ISO, 2007). As stated in (Lee & al., 2004) such representation enables specialists from diverse application fields to share detailed expertise from diverse domains, and implementers to share basic notions which can reduce the overall cost of application development. Feature structures in this description are PoS that apply to Serbian. In this description we have separated those grammatical features and their values that apply to more than one PoS as a supertype in order to avoid redundancy. These descriptions are connected with the appropriate morpho-syntactic categories given in the Data Category Registry (ISO, 2009). For example, the declaration for the grammatical gender is:

```
<fDecl name="kGrGen" xml:id="grammaticalGender">
<fDescr>Gender is the grammatical category….</fDescr>
    <vRange>
        <vAlt>
            <symbol value="m"/><!-- masculine -->
            <symbol value="f"/><!-- feminine -->
            <symbol value="n"/><!-- neutre -->
        </vAlt>
    </vRange>
</fDecl>
```

In order to explicitly formulate relations between categories and their values as well as their applicability we have used feature structure constraints with conditional and bi-conditional tests. For instance, feature structure for nouns consists of nine features, five of which are present as attributes in MULTEXT-East: type, grammatical gender, grammatical number, case, animacy. The added features are: binary feature 'multi-word' which is applicable to all PoS, natural gender, natural number and changeability of the number. Besides that we have added for each of the four main grammatical categories a binary category that states whether grammatical category is fixed or not for nouns in general. It should be noted that this is the property of a category as related to the part of speech and not for the category in general. A part of the feature structure declaration for nouns is:

```
<fsDecl type="st_noun" baseTypes="st_categories">
  <fsDescr> a word that can be used to refer to a person or
  place or thing</fsDescr>
    <fDecl name="POS">
        <vRange><symbol value="N"/></vRange>
    </fDecl>
    <fDecl name="Gen">
        <vRange><fs type="kGrGen"/></vRange>
    </fDecl>
    <fDecl name="FGen">
        <vRange><binary value="false"/></vRange>
    </fDecl> <!-- does not inflect in gender -->
```

Two constraints are defined for nouns: one that states that the number 'paukal' exists only for the genitive and the accusative case, and the other that gives the possible value pairs of the grammatical number in the singular and plural. The first of these constraints is:

```
<cond>
  <fs>
    <f name="broj"> <symbol value="w"/>
    </f><!-- if the number is paukal -->
  </fs>
  <then/>
  <fs>
    <f name="padež"> <!--the case can be only -->
            <symbol value="2"/> <!--the genitive -->
            <symbol value="4"/> <!--the accusative -->
    </f>
  </fs>
 </cond>
```

Some other PoS use more constraints in their description: for instance, there are eight constraints for numerals and six for pronouns. Although we have tried to maintain compatibility with MULTEXT-East it does not mean that there is one-to-one correspondence between our features and MULTEXT-East attribute. For instance, in our description for verbs verbal forms are strictly separated from tense, voice and aspect, as usual in LADL morphological dictionaries in general. Constraints specify which forms are used with these verbal features, and which of them are realized as simple or compound. For that reason constraints for verbs are numerous. However, MULTEXT-East attributes and values can be unambiguously derived.

The recommendations presented in TEI Guidelines (TEIP5) correspond to ISO 24610 standard which enabled the validation of SrpFSD. In order to do that we used Roma tool that allows creation of TEI conformant validators in a form of DTD, Relax NG or W3C Schemas. The appropriate web interface [2] allowed us to generate a reduced schemas and documentation. The new schema combines default modules: core tei, header, textstructure, with the iso-fs module which covers feature structures.

5. The Unification of Solutions

As our ultimate aim is to maintain and further develop only one description, we have prepared two XSLT scripts that transform SrpFSD into descriptions used by Unitex subsystems Elag and Multiflex. The following excerpt from one of the scripts transforms SrpFSD into description used by Elag. This excerpt produces from the FSD described in section 4 the inflectional properties of nouns described in subsection 3c.

[2] http://www.tei-c.org/Roma/

```
<xsl:text >complet: &#xa;</xsl:text>
  <xsl:for-each select="//fsdDecl/fsDecl/
fDecl[@rend='number']/vRange/vAlt/symbol">
     <xsl:call-template name="Complet">
        <xsl:with-param name="nb" select="@value"/>
        <xsl:with-param name="cond"
select="//fsConstraints/cond[fs/f/@name='Nb']"/>
        <xsl:with-param name="val" select="//
fsConstraints/cond/fs/f[@name='Nb']/symbol/@value"/>
     </xsl:call-template>
  </xsl:for-each>…
<xsl:template name="Complet" >
 <xsl:param name="nb"/> <xsl:param name="cond"/>
<xsl:param name="val"/>
 <xsl:choose>
       <xsl:when test="$nb = $val">
        <xsl:for-each select=
"$cond/fs/f[@name='Case']/symbol/@value">
              <xsl:call-template name="CompletRow">
               <xsl:with-param name="n"
select="$nb"/><xsl:with-param name="c" select="."/>
              </xsl:call-template>
        </xsl:for-each>
       </xsl:when>
       <xsl:otherwise>
     <xsl:text >&#xa;  &lt;gender&gt;
</xsl:text><xsl:value-of select="$nb"/>
     <xsl:text > &lt;case&gt; &lt;anim&gt;</xsl:text>
       </xsl:otherwise>
     </xsl:choose>
 </xsl:template>
```

6. Future Work

Our description is not yet finalized. For some problems we have not yet found a satisfactory solution, e.g. the change (not inflection) of the gender with the change of the number for nouns.

In a future we intend to test our description in various applications involving Serbian. We also expect to incorporate our description to MULTEXT-East version 4 that will be published recently (Erjavec, 2009). This new version will solve many problems encountered in the previous versions, such as lack of attributes or attribute values, explicit statement of valid combinations of attribute and values. We expect that our description will transform without loss of information in the new version of MULTEXT-East format.

7. References

Courtois, B. and Silberztein, M. (eds.) (1990). Dictionnaires électroniques du français. Langue française 87. Paris: Larousse.

Erjavec, T. (2004) MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*, pp. 1535 - 1538, ELRA, Paris.

Erjavec, T. MULTEXT-east morphosyntactic specifications : towards version 4. V: Garabík, R. (ed). MONDILEX Third Open Workshop, Bratislava, Slovakia, 15-16 April, 2009. *Metalanguage and encoding scheme design for digital lexicography : innovative solutions for lexical entry design in Slavic lexicography: proceedings*. Bratislava: L'. Štúr Institute of Linguistic, Slovak Academy of Sciences, 2009, str. 59-70.

ISO. (2007) ISO 24610-2:2007 Language resource management - Feature Structures – Part 2: Feature System Declaration, ISO/TC 37/SC 4.

ISO. (2009) ISO 12620 Terminology and other language and content resources – Data Categories – Specification of data categories and management of a data category registry for language resources

Kešelj, V., Kešelj, T., and Zlatić, L. (2004). R{j}ecnik.com: English-Serbo-Croatian electronic dictionary. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries* (Geneva, Switzerland, August 29 - 29, 2004). ACL Workshops. ACL, Morristown, NJ, 61-64.

Krstev, C., Vitas, D. and Erjavec, T. (2004) Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian. In *Informatica*, No. 28, pp. 431-436, The Slovene Society Informatika, Ljubljana.

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

Krstev, C. and Vitas, D. (2009) An Effective Methode for Developing a Comprehensive Morphological E-Dicitonary of Compounds, The 28th Conf. on Lexis and Grammar, Bergen, 29th September - 3rd October 2009, In *Arena Romanistica* eds. B. Lamiroy et al, pp. 204-212, University of Bergen, Department of Foreign Languages.

Laporte, E. and Monceaux, A. (1999). "Elimination of lexical ambiguities by grammars. The ELAG system", *Lingvisticae Investigationes* XXII, Amsterdam-Philadelphie : Benjamins, pp. 341-367.

Lee, K., Burnard, L., Romary, L., de la Clergerie, E., Declerck, T., Bauman, S., Bunt, H., Clement, L., Erjavec, T., Roussanalay, A., Roux, C. (2004) "Towards and international standard on feature structures representation",in *Proc. of LREC'04*, pp.373-376.

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

Popović, Z. (2009) Taggers Applied On Texts On Serbian Language, Language Tools And Machine Learning. In *Infotheca*, Vol. X, No. 2, (to appear).

Przepiórkowski, A. and Woliński, M. (2003) A Flexemic Tagset For Polish. In Proc. of the Workshop on Morphological Processing of Slavic Languages : 10th Conference EACL 2003, Budapest, Hungary, April 13th, 2003, eds. T. Erjavec and D. Vitas, pp. 33-40.

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