INTERPRETING BULGARIAN SOUND ALTERNATIONS OF INFLECTIONAL MORPHOLOGY IN DATR

Velislava Stoykova
Institute for Bulgarian Language, Bulgarian Academy of Sciences, 52, Shipchensky proh. str., bl. 17, 1113 Sofia, Bulgaria

Keywords: Knowledge representation, Semantic networks, Sound alternations, Inflectional morphology.

Abstract: The paper presents an approach to interpret sound alternations of Bulgarian language for inflectional morphology of definite article. The DATR language for lexical knowledge presentation is accepted as a framework, and the analysis and examples of semantic network for different part-of-speech are presented. Finally, more general conclusions for formal interpretation of sound alternations for inflectional morphology are defined.

1 INTRODUCTION

Recent developments in AI offer various approaches to knowledge presentation. Thus, natural language processing applications use different techniques to represent and differentiate between phonological, morphological, and syntactic knowledge even all these types represent the inherent language features.

The sound alternations influence the inflectional morphology of almost all part-of-speech of standard Bulgarian language and as a result they form irregular word forms. In fact, we have a rather unsystematically formed variety of regular and irregular sound alternations which is very difficult to be interpreted formally.

2 THE COMPLEXITY OF THE FORMAL INTERPRETATION

The problem of interpreting sound alternations of inflectional morphology is a key problem of any formal presentation of standard Bulgarian language. In our paper we present different types of sound alternations, and analyse the DATR encoding of the feature of definiteness in Bulgarian and the related inflectional morphology interpretation. We propose a formal account of sound alternations using both the architecture of the application – the principle of local and global inheritance and the lexical information presentation scheme. We offer an interpretation of both phonetic and morphological features in one consistent formal representation.

2.1 Types of Sound Alternations

The phonetic alternations in Bulgarian are of various types. They take place in different historical periods of language development, and form different types of irregular word forms. The sound alternations influence both derivational and inflectional morphology. Further we are going to analyse only those alternations which are significant for the inflectional morphology.

The sound alternations in standard Bulgarian are divided into two general types: (i) phonologically based sound alternations, and (ii) historically based sound alternations (Grammar, 1982). The classification is accepted in the academic descriptive grammar works, and it does not concern the formal way the sound alternations could be interpreted. In further examples sound alternations are illustrated with respect to the grammar features of number and definiteness, which in Bulgarian language are inflectional grammar features.

2.2 Phonologically based Sound Alternations

Phonologically based sound alternations concern mostly vocal sounds. Those of them which reflect nominal inflectional morphology are of three types. The first type of phonologically based sound alternations is the so-called epenthesis of [@][1]. The example

[1]Here we use SAM Phonetic Alphabet (SAMPA) to mark sounds and Latin alphabet for the letters. Because of mismatching between the two alphabets some of the Cyril-

Stoykova V..
INTERPRETING BULGARIAN SOUND ALTERNATIONS OF INFLECTIONAL MORPHOLOGY IN DATR.
DOI: 10.5220/0003293804860491
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 486-491
ISBN: 978-989-8425-40-9

is as follows:

```
[ideali"z@m] - [ideali"zm-@t]
*idealism (sg.undef.) - idealism-the (sg.def.)
```

The second type of phonetically based sound alternations is the methathesis of [r@], [l@] : [@r], [@l]. The example is:

```
[gr@b] - [g@"rb-Ove]
*back (sg.) - back-s (pl.)
```

The third type phonologically based sound alternation is the so-called vocal alternation of ja-slide. The alternation takes place when [3] (ja) is in stressed position and stands before non-palatal syllable. Under these conditions the alternative word form has different vocal sound, namely [e]. The ja-slide divides Bulgarian dialects into Eastern and Western, however, in the standard Bulgarian both regular and irregular forms are accepted. The example is as follows:

```
[gr'ax] - [gre"x-Ove]
*sin (sg.) - sin-s (pl.)
```

2.3 Historically based Sound Alternations

Historically based sound alternations are mostly consonant alternations. The first type is the so-called II palatalization. It could be explained as a result of the more general process of palatalising the consonants, which takes place during centuries and influences all Slavic languages. This alternation changes the consonants [g], [k], [x] into [z], [ts], [s] respectively when they precede the vocals [3], [e] or [i]. The examples are:

```
[pO"dlOg] - [pO"dlOz-i]
*subject (sg.) - subject-s (pl.)

[ve"stnik] - [ve"stnits-i]
*newspaper (sg.) - newspaper-s (pl.)

[mOna"x] - [mOna"s-i]
*monk (sg.) - monk-s (pl.)
```

3 THE INFLECTIONAL NATURE OF DEFINITENESS IN BULGARIAN

The standard Bulgarian language does not use cases for syntactic representation but it has very rich inflectional system - both for derivational and for inflectional morphology, and it uses prepositions and a base word form instead a case declination. It is considered to be a language using relatively free word order, so the subject can take every syntactic position in the sentence (including the last one). Another important grammar feature of Bulgarian is the feature of definite article which is an ending morpheme (Grammar, 1983). The fact gives a priority to morphological interpretations of definiteness in spite of syntactic since, and at the level of syntax, the definite article shows the subject (when it is not a proper name).

3.1 The Definite Article - Its Semantics and Morphosyntactic Features

At syntactic level, the definiteness in Bulgarian may express various types of semantic relationships like a case (to show subject), part-of-whole, deixis etc. The definite article can assign an individual, or quantity definiteness, and it has a generic use as well.

3.2 The Formal Morphological Marker of Definiteness

The syntactic function of definiteness in Bulgarian is expressed by a formal morphological marker which is an ending morpheme (Grammar, 1983). It is different for the genders, however, for the masculine gender two types of definite morphemes exist – to determine a defined in a different way entities, which have two phonetic alternations, respectively. For the feminine and for the neuter gender only one definite morpheme exists, respectively. For the plural, two definite morphemes are used depending on the ending vocal of the main plural form. The following part-of-speech in Bulgarian take the definite article: nouns, adjectives, numerals (both cardinals and ordinals), possessive pronouns (the full forms), and reflexive-possessive pronoun (its full form).

The definite morphemes are the same for all part-of-speech, however, in further description we are going to analyze only some general types of rules used for the interpretation of inflectional morphology of definiteness in Bulgarian given in Stoykova (Stoykova, 2002), (Stoykova, 2004a), (Stoykova, 2010). Thus, our task is to present the related architecture of the application which suggests a subsequent algorithm for the rule-based interpretation of the inflectional morphology of definite article.

lic letters are presented by using two Latin letters.

4 THE COMPUTATIONAL MORPHOLOGY APPROACHES

The standard computational approach to both derivational and inflectional morphology is to represent words as a rule-based concatenation of morphemes, and the main task is to construct the relevant rules for their combinations. With respect to the number and types of morphemes, different theories offer different approaches depending on the variations of either stems or suffixes as follows (Fig.1):

(i) Conjugational solution offers an invariant stem and variant suffixes, and

(ii) Variant stem solution offers variant stems and invariant suffix.

Both these approaches are suitable for languages, which use inflection rarely to express syntactic structures, whereas for those using rich inflection some cases where phonological alternations appear both in stem and in concatenating morpheme a "mixed" approach is used to account for the complexity. Also, some complicated cases where both prefixes and suffixes have to be processed require such approach.

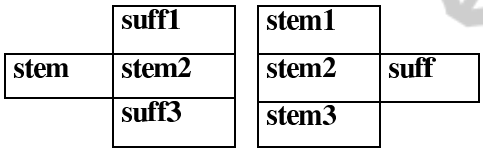

Figure 1: Conjugational solution and variant stem solution.

We evaluate the "mixed" approach as a most appropriate for our task because it considers both stems and suffixes as variables and, also, can account for the specific phonetic alternations. The additional requirement is that during the process of the inflection all generated inflected rules (both using prefixes and suffixes) have to produce more than one type of inflected forms.

Thus, the DATR language for lexical knowledge presentation is a suitable formal framework for presenting inflectional morphology of Bulgarian definite article.

5 THE DATR LANGUAGE

The DATR language for lexical knowledge presentation offers syntagmatic operators (which can be used to define the concatenation procedures) and paradigmatic operators (which can be used to define the specific structure of the inflecting rules, and for further development of the part-of-speech interpretations).

The DATR language is a non-monotonic language for defining the inheritance networks through path/value equations (Evans and Gazdar, 1996). It has both an explicit declarative semantics and an explicit theory of inference allowing efficient implementation, and at the same time, it has the necessary expressive power to encode the lexical entries presupposed by the work in the unification grammar tradition (Evans and Gazdar, 1989a), (Evans and Gazdar, 1989b).

In DATR information is organized as a network of nodes, where a node is a collection of related information. Each node has associated with it a set of equations that define partial functions from paths to values where paths and values are both sequences of atoms. Atoms in paths are sometimes referred to as attributes.

DATR is functional, it defines a mapping which assigns unique values to node attribute-path pair, and the recovery of this values is deterministic. With respect to its universality, DATR's formal properties and techniques underlay both the rule-based inference and non-monotonic inference by default, and allow to account for language phenomena such as regularity, irregularity, and subregularity, by using deterministic parsing.

The semantics of DATR uses non-monotonic inference and default inheritance, and allows the generalization-capturing representation of the inflectional morphology. DATR has the expressive power which is capable to encode and process both syntactic and morphological rules and it allows representation of grammar knowledge by using the semantic networks.

The DATR language has a lot of implementations, however, the analysed application was made by using QDATR 2.0 (`consult http://www.cogs.susx.ac.uk/lab/nlp/datr /datrnode49.html` for a related file `bul_det.dtr`. This PROLOG encoding uses Sussex DATR notation (DATR, 1997). DATR allows construction of various types of language models (language theories), however, the analysed application presents a rule-based formal grammar and a lexical database. The particular query to be evaluated is related inflected word forms, and the implementation allows to process words in Cyrillic alphabet.

6 THE ARCHITECTURE OF THE MODEL

The analyzed application of inflectional morphology of Bulgarian definite article is linguistically motivated. In particular, the underlying basic idea is that

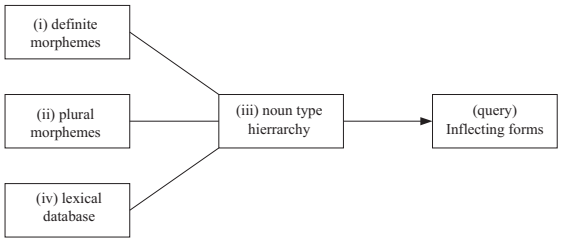

Figure 2: The general architecture of the model.

of a paradigm since morphemes are defined to be of semantic value and are considered as a realisation of a specific morphosyntactic phenomenon. The words are encoded by introducing different roots to account for the related phonetic alternations, which are defined to be of semantic value as well.

Some other DATR applications which present Slavonic inflectional morphology are available for Polish, Russian, Czech, Slovene. The nominal inflectional morphology interpretations deals mostly with case morphology presentation (Czuba, 1994), (Corbett and Fraser, 1993), however they do not offer a systematic account of sound alternations. As for verbal inflectional morphology presentation, the interpretations (Skoumalova, 1996), (Brown, 1998), (Erjavec, 1992) use as underlying idea that of a conjugation.

The architecture represents an inheritance network consisting of various nodes which allows to account for all related inflected forms within the framework of one grammar theory. Thus, the general architecture of the application is as follows (Fig. 2):

(i) All definite inflecting morphemes for all forms of definite article attached to node `DET` and defined by their values by the paths `<masc>`, `<masc_1>`, `<femn>`, `<neut>`, and `<plur>`.

(ii) 12 inflecting morphemes for generating plural forms defined at node `Suff`.

(iii) The inflectional rules defined as concatenations of morphemes for generation of all possible inflected forms attached to the related inflectional types nodes.

(iv) The words are given as lexical database attached to their inflectional type nodes. They are defined as lexical entries through paths `<root>`, `<root gend>`, and `<root plur>`, so to account for the different phonological alternations.

The DATR logical representation framework uses rule-based reasoning with non-monotonic inference and default inheritance to represent the inflectional rules in semantic network. It suggests the structure of semantic network that can employ the generalization-capturing rules in which the grammar knowledge is encoded by the attachment of inflectional rules to the related nodes. In principle, DATR permits multiple default inheritance and prioritized inheritance enforced by orthogonal representation, and suggest the lexicon being structured mostly by inheritance. This technique allows to account for the grammar irregularities and to use the compilation rules which can generate all possible inflected forms within one application.

pref1	pref2	root	suff1	suff2	ending

Figure 3: The word structure according to the general linguistic morphological theory.

The general morphological theory offers a segmentation of words (Fig. 3) which consists of root to which prefixes, suffixes or endings are attached. In Bulgarian, all three types of morphemes are used and a hierarchical structure of the lexical representation in which the feature of gender is a trigger to change the values of the inflected forms is used. During the process of inflection, also, various phonetic alternations are taking place. The phonetic alternations at the morpheme boundary are interpreted either by defining new grammar rules or new nodes, and the phonetic alternations inside morphemes are interpreted by introducing different roots. It is possible, also, to use the technique of finite state transducers (Stoykova, 2004b).

The analysed application interprets, also, more complicated cases of inflection, where both prefixes and suffixes can be processed by defining new nodes of the network.

7 THE INFLECTIONAL RULES

The DATR analysis of nouns (Stoykova, 2002) starts with node `DET` which defines all inflecting morphemes for the definite article and is as follows[2]:

```
DET:
     <sing undef>        ==
     <sing def_2 masc>   == _ja
     <sing def_2 masc_1> == _a
     <sing def_1 masc>   == _jat
     <sing def_1 masc_1> == _ut
     <sing def_1 femn>   == _ta
     <sing def_1 neut>   == _to
```

[2]Here and elsewhere in the description we use Latin alphabet to present morphemes instead of Cyrillic used normally. Because of mismatching between both some of typically Bulgarian phonological alternations are assigned by two letters, whereas in Cyrillic alphabet they are marked by one.

```
    <plur undef>        ==
    <plur def_1>        == _te.
```

The node Suff defines 12 inflecting morphemes for generating the plural inflected forms.

```
Suff:
    <suff_11>  == _i
    <suff_111> == _ovci
    <suff_12>  == _e
    <suff_121> == _ove
    <suff_122> == _eve
    <suff_123> == _ovce
    <suff_21>  == _a
    <suff_22>  == _ja
    <suff_211> == _ishta
    <suff_212> == _ta
    <suff_213> == _ena
    <suff_214> == _esa.
```

The basic node of the nouns inflectional types is the node Noun.

```
Noun:
  <suff> == suff_11
  <gender> == masc_1
  <> == <stem> DET: <Idem "<gender>">
  <stem sing> == "<root sing>"
  <stem plur>=="<root plur>"Suff:<"<suff>">.
```

The example lexeme of word for traveller 'putnik' is defined by <root> and <root plur>.

```
Putnik: <> == Noun
      <root> == putnik
      <root plur> == putnic.
```

The following inflected forms are generated according to the defined rules:

```
Putnik: <gender> == masc_1.
Putnik: <sing undef> == putnik.

Putnik: <plur undef> == putnic_i.
Putnik: <sing def_1> == putnik_ut.
Putnik: <sing def_2> == putnik a.
Putnik: <plur def_1> == putnic_i_te.
```

The adjectives inflectional types hierarchy (Stoykova, 2004a) uses definite morphemes of node DET and the basic node of the hierarchy defines the grammar rules for generating all inflected forms for the feature of gender, number, and definiteness as follows.

```
AdjG:
 <sing undef masc> == "<root>"
 <sing undef femn> == "<root gend>" _a
 <sing undef neut> == "<root gend>" _o
 <sing def_2 masc> =="<plur undef masc>" DET
 <sing def_1 masc> =="<plur undef masc>" DET
 <sing def_1>      == "<sing undef>" DET
 <plur undef>      == "<root gend>" _i
 <plur def_1>      == "<plur undef>" DET.
```

Node Adj inherits all grammar rules of node AdjG but employs also the grammar rules for generating forms for comparison of degree which are produced by using prepositional morphemes.

```
Adj:
     <> == AdjG
     <compar> == po_  "<>"
     <superl> == naj_ "<>".
```

Node Adj_1 inherits the grammar rules of node Adj but changes the rules for generating singular forms for feminine and for neuter gender to use the palatal gender morphemes.

```
Adj_1:
<> == Adj
<sing undef femn> == "<root gend>" _ja
<sing undef neut> == "<root gend>" _jo.
```

Node Adj_2 defines the inflectional rules of the adjectives, which realise two types of phonetic alternations during the process of inflection. At this node additional inflectional base form <root plur> is introduced to account for the complexity.

```
Adj_2:
     <> == Adj
    <plur undef> == "<root plur>" _i.
```

The example lexeme for word short 'tesen' is defined as follows (the generated inflected forms are given at the Appendix):

```
Tesen: <> == Adj_2
     <root> == tesen
     <root gend> == tjasn
       <root plur> == tesn.
```

The reflexive-possessive pronoun uses inflectional rules (Stoykova, 2010) as follows:

```
Adj_5:
<> == Adj
<sing undef femn> == "<root gend>" _ja
<sing undef neut> == "<root gend>" _e
<sing def_1 masc> == "<root gend>" DET
<sing def_2 masc> == "<root gend>" DET.
```

The example lexeme of the reflexive-possessive pronoun self 'svoj' is defined as follows (the generated inflected forms are given at the Appendix):

```
Svoj:  <> == Adj_5
       <root> == svoj
       <root gend> == svo.
```

8 CONCLUSIONS

The proposed analysis uses linguistic motivation for the encoding. The structure of semantic hierarchy uses the grammar feature with a wider scope as a trigger to change the value of inflected forms. The interpretation do not offer a systematic account of sound alternations but suggests some ideas for presenting them by using both architecture of the application – the principle of local and global inheritance, and lexical information presentation scheme. Most of inflectional type nodes are defined to account for

different phonological alternations. The phonetic alternations at the morpheme boundary are interpreted either by defining new grammar rules or new nodes, and the phonetic alternations inside morphemes are interpreted by introducing different roots.

The analyzed application of Bulgarian inflectional morphology accounts for orthographic principles, phonetic alternations, and morphological dependencies and is expected to be reencoded and implemented in lexicography for producing different types of dictionaries.

REFERENCES

Brown, D. (1998). Stem indexing and morphophonological selection in the russian verb. In *R. Fabri, A. Ortmann, and T. Parodi (eds.) Models of Inflection*. Tuebingen: Niemeyer, 196-221.

Corbett, G. and Fraser, N. (1993). Network morphology: a datr account of russian nominal inflection. In *Journal of Linguistics*. 29, 113-142.

Czuba, K. (1994). The datr web pages at sussex. In *http://www.cogs.susx.ac.uk/lab/nlp/datr/ datrnode49, file polish_n.dtr*

DATR (1997). The datr web pages at sussex. In *http://www.cogs.susx.ac.uk/lab/nlp/datr/*.

Erjavec, T. (1992). Treatments of slovene verb morphology in inheritance models. In *MSc Thesis*. Edinburgh.

Evans, R. and Gazdar, G. (1989a). Inference in datr. In *Fourth Conference of the ECACL*. 66-71.

Evans, R. and Gazdar, G. (1989b). The semantics of datr. In *Anthony G. Cohn, ed. Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*. 79–87.

Evans, R. and Gazdar, G. (1996). Datr: A language for lexical knowledge representation. In *Computational Linguistics*. 22(2):167–216.

Grammar (1982). Gramatika na suvremennia bulgarski knizoven ezik. In *Fonetika, tom. 1*.

Grammar (1983). Gramatika na suvremennia bulgarski knizoven ezik. In *Morphologia, tom. 2.*

Skoumalova, H. (1996). Czech hierarchical lexicon. In *http://utkl.ff.cuni.cz/∼skoumal/czech_verb.html.*

Stoykova, V. (2002). Bulgarian noun – definite article in datr. In *D. Scott, ed. Artificial Intelligence: Methodology, Systems, and Applications. LNAI 2443*. Springer, 152–161.

Stoykova, V. (2004a). The definite article of bulgarian adjectives and numerals in datr. In *C. Bussler and D. Fensel, eds. Artificial Intelligence: Methodology, Systems, and Applications. LNAI 3192*. Springer, 256–266.

Stoykova, V. (2004b). Modeling sound alternations of bulgarian language in datr. In *E. Buchberger ed. Proceedings of KONVENS 2004, Shriftenreiche der Oesterreichischen Gesellschaft fur Artificial Intelligence*. Band 5, Wien, 201-204.

Stoykova, V. (2010). Bulgarian possessive and reflexive-possessive pronouns in datr. In *R. Trappl ed. Cybernetics and Systems 2010*. 426-432.

APPENDIX

```
Tesen: <sing undef masc> == tesen.
Tesen: <sing undef femn> == tjasna.
Tesen: <sing undef neut> == tjasno.
Tesen: <plur undef> == tesni.
Tesen: <compar sing undef masc> == po-tesen.
Tesen:<compar sing undef femn> == po-tjasna.
Tesen:<compar sing undef neut> == po-tjasno.
Tesen: <compar plur undef> == po-tesni.
Tesen:<superl sing undef masc> == naj-tesen.
Tesen:<superl sing undef femn>== naj-tjasna.
Tesen:<superl sing undef neut>== naj-tjasno.
Tesen: <superl plur undef> == naj-tesni.
Tesen: <sing def_1 masc> == tesnijat.
Tesen: <sing def_2 masc> == tesnija.
Tesen: <sing def_1 femn> == tjasnata.
Tesen: <sing def_1 neut> == tjasnoto.
Tesen: <plur def_1> == tesnite.
Tesen: <compar sing def_1 masc>==po-tesnijat.
Tesen: <compar sing def_2 masc>==po-tesnija.
Tesen: <compar sing def_1 femn>==po-tjasnata.
Tesen: <compar sing def_1 neut>==po-tjasnoto.
Tesen: <compar plur def_1> == po-tesnite.
Tesen:<superl sing def_1 masc>==naj-tesnijat.
Tesen: <superl sing def_2 masc>==naj-tesnija.
Tesen:<superl sing def_1 femn>==naj-tjasnata.
Tesen:<superl sing def_1 neut>==naj-tjasnoto.

Svoj: <sing undef masc> == svoj.
Svoj: <sing undef femn> == svoja.
Svoj: <sing undef neut> == svoe.
Svoj: <plur undef> == svoi.
Svoj: <sing def_1 masc> == svojat.
Svoj: <sing def_2 masc> == svoja.
Svoj: <sing def_1 femn> == svojata.
Svoj: <sing def_1 neut> == svoeto.
Svoj: <plur def_1> == svoite.
```