M a t h e m a t i c a
B a l k a n i c a

New Series Vol. 26, 2012, Fasc. 3–4

The Geometry of Language – a Space Semantic Network of Bulgarian Nominal Inflectional Morphology *

Velislava Stoykova[1], *Chavdar Lozanov*[2]

Presented at MASSEE International Conference on Mathematics MICOM-2009

In this work we present a computationally tractable model of Bulgarian nominal inflectional morphology for the feature of definiteness using orthogonal semantic networks. We use the DATR language for lexical knowledge presentation for the implementation, and we define the basic architecture of the model in terms of its linguistic motivation, namely, the semantics and the grammar features of definiteness in Bulgarian language. Further, we explain a fragment of semantic network encoding of nouns, and evaluate different inflected forms. Finally, we offer a geometrical interpretation of the encoded semantic network in the space.

MSC 2010: 68T50, 68T30.

Key Words: Bulgarian nominal inflectional morphology, computational model, DATR language.

Introduction

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 noun 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 [5]. 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

*Partially supported by Grant No 130/2009 of Science Fund of University of Sofia "St. Kliment Ohridsky"

proper name). So that, modeling inflectional morphology of definite article is important stage of a successful part-of-speech parsing of Bulgarian.

The semantics of definiteness and its formal morphological marker

According to traditional academic descriptive grammar works [5], the semantics of definiteness in standard Bulgarian is expressed in three ways: lexical, morphological, and syntactic. The lexical way is closely related to lexical semantics of a particular lexeme. At syntactic level, the definiteness in Bulgarian express various types of semantic relationships like a case (to show subject), part-of-whole, deixis, etc. Also, the definite article can assign an individual or quantity definiteness, and it has a generic use as well.

The syntactic function of definiteness in Bulgarian is expressed by a formal morphological marker which is an ending morpheme. It is different for genders, however, for masculine gender two types of definite morphemes exist – to determine entities defined in a different way, which have two phonetic alternations, respectively. For feminine and for neuter gender only one definite morpheme exists, respectively. For plural, two definite morphemes are used depending on the ending vocal of the main plural form. The following part-of- speech in Bulgarian take definite article: nouns, adjectives, numerals (both cardinals and ordinals), possessive pronouns (the full forms), and reflexive-possessive pronoun (its full form). The definite article is the same for all part-of-speech but it has different forms to account for the feature of gender and number. Further, we are going to analyze the interpretation of the nominal inflectional morphology of definiteness in Bulgarian given in Stoykova [7].

The traditional academic representation and computational morphology formal models representation of inflectional morphology

The traditional interpretation of inflectional morphology given at the academic descriptive grammar works [5] is a presentation of tables. The tables consist of all possible inflected forms of a related word with respect to its subsequent grammar features. The artificial intelligence (AI) techniques offer a computationally tractable encoding preceeded by a related semantic analysis, which suggest a subsequent architecture. Reprsenting inflectional morphology in AI frameworks is, in fact, to represent a specific type of grammar knowledge.

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 relevant rules for their combinations. With respect to number and types of morphemes, the different theories offer

different approaches depending on variations of either stems or suffixes as follows: (i) Conjugational solution offers invariant stem and variant suffixes, and (ii) Variant stem solution offers variant stems and invariant suffix. However, for Bulgarian we evaluate the "mixed" approach as a most appropriate because it considers both stems and suffixes as variables and, also, can account for the specific phonetic alternations. Also, we use DATR language for lexical knowledge presentation as a suitable formal framework for presenting inflectional morphology of Bulgarian definite article.

The DATR language

The DATR language is a non-monotonic language for defining inheritance networks through path/value equations [4]. 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 [2, 3, 4]. 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. It can account for such language phenomena like regularity, irregularity, and subregularity, and allows the use of deterministic parsing. The DATR language has a lot of implementations, however, our application was made by using QDATR 2.0 (consult `URL 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 [8]. DATR allows construction of various types of language models (language theories), however, our model is presented as a rule-based formal grammar and a lexical database. The particular query to be evaluated is a related inflecting word form, and the implementation allows to process words in Cirillic alphabet.

The principles and architecture of DATR encoding of Bulgarian nominal inflectional morphology

Our model is linguistically motivated and presents a rule based formal grammar and a lexical database. We define morphemes to be of semantic value and consider them as a realization of a specific morpho-syntactic phenomena. We are encoding words using the traditional notion of the lexeme, and we accept different sound alternations to be of semantic value. The model represents an inheritance network consisting of various nodes, and allows us to account for all

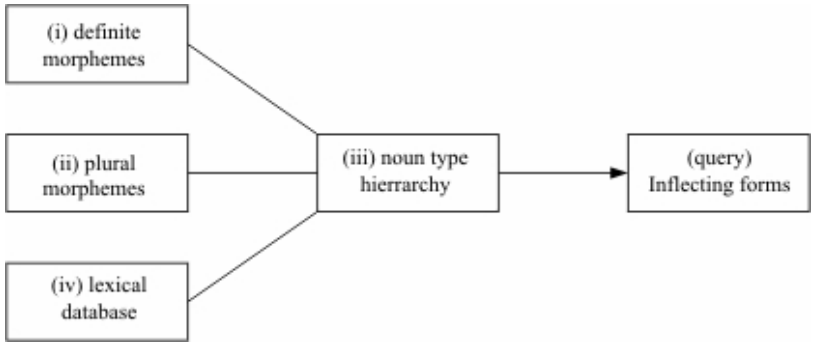

Figure 1: The overall architecture of Bulgarian noun inflection

related inflected forms. The approach developed is indebted to that of Cahill and Gazdar [1] used to account for German noun inflection. The general architecture of the application is as follows (see Fig. 1.):

(i) All definite inflecting morphemes for all forms of definite article attached to node `DET` and defined by their values through paths `<masc>`, `<masc_1>`, `<femn>`, and `<neut>`. (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 a lexical database attached to their inflectional type nodes, respectively. They are defined by the lexical entries through paths `<root>`, and `<root plur>`, so to account for the different phonological alternations. The noninflectional features are given as invariables, and are defined with their particular values for the related words. (v) The queries to be evaluated are all possible inflected word forms which are produced after the stage of compilation.

A fragment of DATR nominal inflection encoding

The Bulgarian noun has three grammar features: gender, number, and definiteness. Among them only number and definiteness are inflectional whereas gender is not inflectional. Thus, we consider gender as a specific trigger in our formal interpretation. With respect to gender, nouns are divided into three groups: of masculine, of feminine, and of neuter gender. Within groups, there are different types of nouns depending on their suffix for forming plural, so that, the next triggering factor is the feature of number. The detailed DATR encoding of all nominal inflection types is given in Stoykova [7]. The table of Bulgarian nominal inflection type hierarchy is presented in Fig. 2. Further, we will analyze

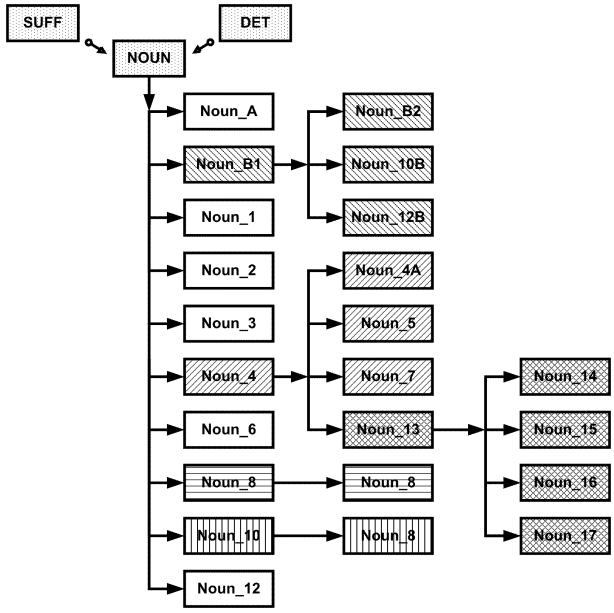

Figure 2: The table presentation of noun inflectional type hierarchy

only a fragment of the encoding, and we start with node `DET` which defines all inflecting morphemes for the definite article 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
         <plur undef>  ==
         <plur def_1> == _te.
```

Also, we define node `Suff` as consisting of 12 ending morphemes for plural.

```
Suff:    <suff_11> == _i
         <suff_111> == _ovci
```

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

```
<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.
```

Node `Noun` is a basic in the semantic network. It takes the information given through paths `<root>` and `<root plur>` for the `<stem>` (the morphemes of a related lemma), and `<gender>` (for the inflected morphemes of a related gender defined at node `DET`), and `<plur>` (for the related plural inflected morphemes defined at node `Suff`). The node consists of grammar rules for generating all inflected forms for the features of number and definiteness.

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

All subsequent inflectional type nodes inherit the grammar rules of node `Noun` and simply change the value of the inflecting morpheme for either number or gender as node `Noun_1`, which changes value for the plural morpheme.

```
Noun_1: <> == Noun
        <suff> == suff_12.
```

Alternatively, node `Noun_10` changes the value of the inflecting morpheme for the gender.

```
Noun_10: <> == Noun
         <gender> == femn.
```

Evaluating different inflected forms

An example lexeme for bulgarian word for *newspaper (vestnik)*, which uses inflectional rules defined at node `Noun` is as follows:

```
Vestnik: <> == Noun
         <root> == vestnik
         <root plur> == vestnic.
```

Following the consequence of given axioms, we can generate the queries which are all possible inflected forms.

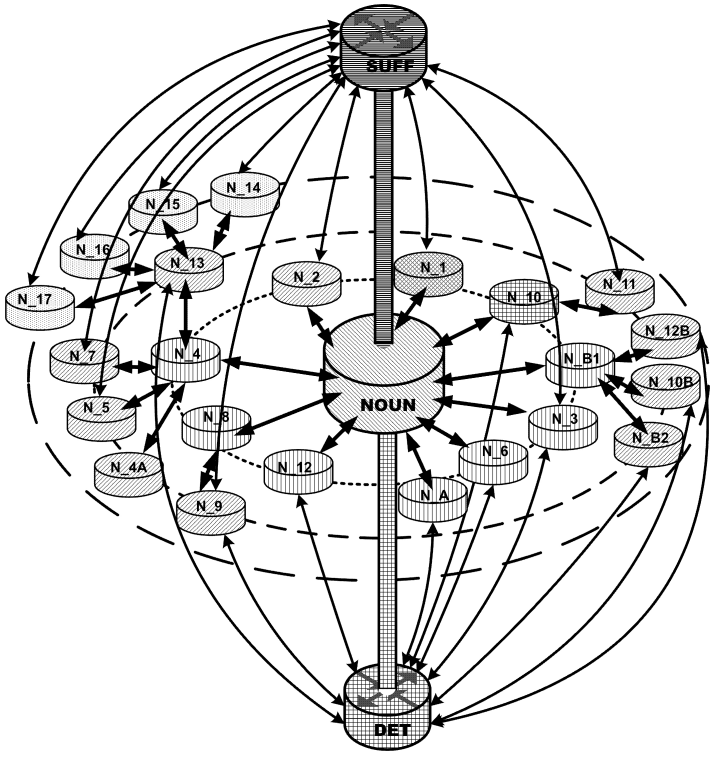

Figure 3: The space representation of noun inflectional type hierarchy.

```
Vestnik: <gender> == masc_1.
Vestnik: <sing undef> == vestnik.
Vestnik: <plur undef> == vestnic_i.
Vestnik: <sing def_1> == vestnik_ut.
Vestnik: <sing def_2> == vestnik_a.
Vestnik: <plur def_1> == vestnic_i_te.
```

The geometrical interpretation of DATR nominal inflection encoding

Our model represents grammar knowledge using orthogonal semantic networks which alows us to offer a geometrical interpretation (see Fig. 3.).

Normally, the visualisation of semantic networks techniques is based on the convertion of orthogonality relation which is interpreted as a semantic relationship [6].

We use such idea, and in our space representation, the geometric coordinates of inflectional nodes underlie the semantic representation of the encoding. Thus, our space model interpret inheritance relationship as a semantic (using three concentric circles in a plane) to present the three types of inflectional nodes. The presentation can be interpreted with respect to all semantic relationships between nodes (including inflection, inheritance, gender or number).

References

[1] L. Cahill, G. Gazdar. German noun inflection. // *Journal of Linguistics*, **35.1**, 1999, 1–42.

[2] R. Evans, G. Gazdar. Inference in DATR. // *Fourth Conference of the European Chapter of the Association for Computational Linguistics*, 1989, 66–71.

[3] R. Evans, G. Gazdar. The semantics of DATR. // Anthony G. Cohn (ed.) *Proceedings of the Seventh Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour.* London, 1989, 79–87.

[4] R. Evans, G. Gazdar. DATR: A language for lexical knowledge representation. // *Computational Linguistics*, **22.2**, 167–216.

[5] *Grammar of the Modern Bulgarian Laguage.* Vol. 2, Morphology. Sofia, 1983 (in Bulgarian).

[6] M. Mesina, D. Roller, C. Lampasona. Visualisation of Semantic Networks and Ontologies Using AutoCAD. // Y. Luo (ed.). *Cooperative Design, Visualisation, and Engineering*, Lecture Notes in Computer Science 3190, Springer-Verlag, 2004, 21–29.

[7] V. Stoykova. Bulgarian noun – definite article in DATR. // D. Scott (ed.). *Artificial Intelligence: Methodology, Systems, and Applications.* Lecture Notes in Artificial Intelligence 2443, Springer-Verlag, 2002, 152–161.

[8] The DATR Web Pages at Sussex URL
`http://www.cogs.susx.ac.uk/lab/nlp/datr/datrnode49.html`

[1] *Institute for Bulgarian Language*
Bulgarian Academy of Sciences
52, Shipchensky proh. str., bl. 17
Sofia 1113, BULGARIA
E-MAIL: vili1@bas.bg

[2] *Department of Geometry*
University of Sofia
5, J. Bourchier blv.
Sofia 1164, BULGARIA
E-MAIL: lozanov@fmi.uni-sofia.bg

Received 04.02.2010