Representing Lexical Knowledge for Bulgarian Inflectional Morphology in DATR

VELISLAVA STOYKOVA
Institute for Bulgarian Language
Bulgarian Academy of Sciences
52, Shipchensky proh. str. bl. 17, 1113 Sofia
BULGARIA
vili1@bas.bg

Abstract: The paper analyses the application of DATR language for lexical knowledge presentation for interpreting Bulgarian inflectional morphology. It discusses the semantic network of the feature of definiteness in Bulgarian language and compares the lexical knowledge representation for the different part-of-speech with respect to the defined grammar rules, the sound alternations, the related formal presentation – both table and computational – and its application to lexicography.

Key–Words: Grammar knowledge presentation, semantic networks, computational morphology, DATR language for lexical knowledge presentation, semantic networks visualisation.

1 Introduction

Recent development in the information technologies offers artificial intelligence based methodologies for almost all fields of scientific research, and can result in a wide range of real applications. These applications can significantly improve the effectiveness of the research especially in such part of it where a researcher needs to perform tasks in searching large-scale database, and using his own knowledge and intuition to make decisions about them. The modern lexicography uses extensively the electronic text corpora of different genres instead text archives used in the past. Existing electronic text collections of different types (www-source, electronic archives, text corpora, etc.) increase the necessity of proper software tools assisting different types of search to make the best use of them. Different dictionaries offer different conception of word definition of the sense but the problem of optimizing searching and finding the word context is central. It is specific for flective languages where different inflected forms share a common meaning. So, the problem of creating tools for knowledge-based search in large-scale electronic text corpora able to deal with both base and inflected word form is a central for optimizing the process of compilation of different types of dictionaries. Further, we are going to discuss the principles and application of a computationally tractable model of Bulgarian inflectional morphology.

2 Bulgarian language is a flective language

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 [11]. The fact gives a priority to morphological interpretations of definiteness in spite of syntactic. At the level of syntax, the definite article shows the subject (when it is not a proper name). So that, modeling inflectional morphology of definite article is important stage of a successful part-of-speech parsing of Bulgarian.

3 The semantics of definiteness and its formal morphological marker

According to traditional academic descriptive grammar works [11], 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 interpretations of definiteness in Bulgarian given in Stoykova [13] for nominal inflectional morphology, Stoykova [14] for adjectives and numerals inflectional morphology, and Stoykova [17] for pronouns inflectional morphology, and mainly the principles and types of lexical knowledge representation of these applications.

4 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 [11] 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 preceded by a related semantic analysis, which suggest a subsequent architecture. Representing 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 a “mixed” approach is the most appropriate because it considers both stems and suffixes as variables and, also, can account for the specific phonetic alternations. Also, 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 is a non-monotonic language for defining inheritance networks through path/value equations [10]. 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 [7], [8], [9]. 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, the analyzed 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 [18]. DATR allows construction of various types of language models (language theories), however, the entire application presents a rule-based formal grammar and a lexical database. The particular query to be evaluated is a related inflected word forms, and the implementation allows to process words in Cyrillic alphabet.

6 The general architecture of the application

The analyzed application of inflectional morphology of Bulgarian definite article is linguistically motivated. In particular, the underlying basic idea is that 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 following the traditional notion of lexeme and different roots are introduced to account for the related phonetic alternations, which are defined to be of semantic value. Some other DATR applications which present Slavonic inflectional morphology are available for Polish, Russian, Czech, Slovene. They offer different insights for presenting both nominal inflectional morphology and verbal inflectional morphology. The nominal inflectional morphology interpretations deals mostly with case morphology presentation and use as underlying idea that of a paradigm – like the applications made for Polish [5] and Russian [4]. As for the verbal inflectional morphology presentation, the interpretations for Czech [12], for Russian [1], and for Slovene [6] use as underlying idea that of a conjugation. Some ideas for inflectional morphology representation are indebted to that of Cahill and Gazdar [2], [3] used to account for German inflectional morphology, however, their account of morphotactics was not applied. The architecture represents an inheritance network consisting of various nodes which allows to account for all related inflected word forms within the framework of one grammar theory. Thus, the general architecture of the application is as follows (Fig. 1):

(i) All definite inflecting morphemes for all forms of definite article attached to node DET and defined by their values through the paths <masc>, <masc_1>, <femm>, <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 a lexical database attached to their inflectional type nodes, respectively. They are defined by the lexical entries through paths <root>, <root gend>, and <root plur>, so to account for the different phonological alternations. The non-inflectional features are given as invariables, and are defined with their particular values for the related words (only for the pronouns).

(v) The queries to be evaluated are all possible inflected word forms which are produced after the stage of the compilation.

7 The principles of lexical knowledge encoding for Bulgarian inflectional morphology

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. The general morphological theory offer a segmentation of word which consists of root to which prefixes, suffixes or endings are attached. In Bulgarian, all three types of morphemes are presented. However, different morphemes are used to account for the feature of gender and number, which suggest the hierarchical structure of the lexical representation in which the feature of gender is a trigger to change the values of the inflected forms. 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 [15]. The application [16] interprets, also, more complicated cases of inflection, where both prefixes and suffixes can be processed by defining different nodes of the network. The semantic principle is used, also, for the encoding of nouns, adjectives, numerals, and pronouns since the inflectional rules are represented by taking into account the semantics of the grammar features of the related part-of-speech and their internal hierarchy. In general, the interpretation presents a tabular conceptualisation inference task of which the conceptual representation is based on accounting for the orthographic, phonetic, morphological, and semantic properties.
7.1 Inflectional rules and lexical representation examples

Thus, we will analyze the fragments of DATR encoding [13], [14], [17] and we start the analysis of nouns inflectional morphology with node DET which defines all inflecting morphemes for the definite article for all part-of-speech as follows:

\[ \text{DET: } \text{sing undef} == \text{sing def_2 masc} == \text{ja} \]
\[ \text{sing def_2 masc_1} == \text{a} \]
\[ \text{sing def_1 masc} == \text{ja} \]
\[ \text{sing def_1 masc_1} == \text{ut} \]
\[ \text{sing def_1 femn} == \text{ta} \]
\[ \text{sing def_1 neut} == \text{to} \]
\[ \text{plur undef} == \text{plur def_1} == \text{te}. \]

Node Suff defines all inflected morphemes for plural and is as follows.

\[ \text{Suff: } \text{suff_11} == \text{i} \]
\[ \text{suff_111} == \text{ovci} \]
\[ \text{suff_12} == \text{e} \]
\[ \text{suff_121} == \text{ove} \]
\[ \text{suff_122} == \text{eve} \]
\[ \text{suff_123} == \text{ovce} \]
\[ \text{suff_21} == \text{a} \]
\[ \text{suff_22} == \text{ja} \]
\[ \text{suff_211} == \text{ishta} \]
\[ \text{suff_212} == \text{ta} \]
\[ \text{suff_213} == \text{ena} \]
\[ \text{suff_214} == \text{esa}. \]

Node Noun defines the inflectional rules for nouns. It takes the information given through paths <root> and <root plur> (defined at the related lexeme), for the <stem>, 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 noun word forms for the features of number and definiteness.

\[ \text{Noun: } <\text{suff}> == \text{suff_11} \]
\[ <\text{gender}> == \text{masc}_1 \]
\[ <> == <\text{stem}> \text{DET:<Idem }"\text{gender}>" \]
\[ <\text{stem sing}>=="<\text{root sing}>" \]
\[ <\text{stem plur}>=="<\text{root plur}>"<\text{Suff:}"\text{suff}>. \]

\[ ^1 \text{Here and elsewhere in the description we use Latin alphabet to present morphemes instead Cyrillic used normally. Because of mismatching between both some of typically Bulgarian phonological alternations are assigned by two letters, whereas in Cyrillic alphabet they are marked by one.} \]
The reflexive-possessive pronouns use adjectives inflectional rules [17]. They have in addition the agreement non-inflectional feature of gender and number and the feature of person, which are encoded as a lexical information and are defined at the related lexemes as invariables. The example lexeme for the word his ‘negov’ is as follows:

Negov:
<> == Adj
<person> == third
<number> == sing
<gender> == masc
<root> == negov
<root gend> == negov.

The generated inflected forms are as follows:

Negov: <sing undef masc> == negov.
Negov: <sing undef femn> == negova.
Negov: <sing undef neut> == negovo.
Negov: <plur undef> == negovi.
Negov: <sing def_1 masc> == negovijat.
Negov: <sing def_2 masc> == negovija.
Negov: <sing def_1 femn> == negovata.
Negov: <sing def_1 neut> == negovoto.
Negov: <plur def_1> == negovite.

8 Conclusion

The analyzed application of Bulgarian inflectional morphology encodes the lexical information using hierarchical linguistically motivated representation based on the traditional notion of lexeme. It is a tabular conceptualization inference presentation which accounts for the orthography principles, phonetic alternations, and morphological dependencies. The model can be used in lexicography for different types of context search since different word forms share the same meaning. It can be useful, also, for automatic compilation of orthographic dictionaries. Further, it would be useful to offer a syntactic interpretation.

References: