

Bulgarian Inflectional Morphology in Universal Networking Language

Velislava STOYKOVA

INSTITUTE FOR BULGARIAN LANGUAGE - BAS, 52, Shipchensky proh. str., bl. 17, 1113 Sofia, Bulgaria
vstoykova@yahoo.com

ABSTRACT

The paper presents a web-based application of semantic networks to model Bulgarian inflectional morphology. It demonstrates the general ideas, principles, and problems of inflectional grammar knowledge representation used for encoding Bulgarian inflectional morphology in Universal Networking Language (UNL). The analysis of UNL formalism is outlined in terms of its expressive power to present inflection, and the principles and related programming encodings are explained and demonstrated.

KEYWORDS: Morphology and POS tagging, Grammar and formalisms, Underresourced languages.

1 Introduction

Modeling inflectional morphology is a key problem for any natural language processing application of Bulgarian language. It can result in a wide range of real applications however different formal models and theories offer different insights for encoding of almost all grammar features, and allow the use of related principles for encoding.

2 General problems with applications of word inflectional morphology

The problems with natural language processing applications for word inflectional morphology are generally of two types (i) the problems of language theory at the level of phonology, morphology, and morphology, and (ii) the adequacy of existing methodologies and techniques to offer the applications capable to interpret the complexity of natural language phenomena. Thus, the context of natural language formal representations and interpretations of inflectional morphology is the logical framework which are capable to deal with regularity, irregularity, and subregularity and have to provide a logical basis for interpreting such language phenomena like suppletion, syncretism, declension, conjugation, and paradigm.

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

The traditional interpretation of inflectional morphology given at the academic descriptive grammar works (Popov and Penchev, 1983) 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 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. The problem how to segment words into morphemes is central and there are two basic approaches of interpretation (Blevins, 2001). The first is Word and Paradigme (WP) approach which uses paradigme to segment morphemes. The second is Item and Agreement (IA) approach which uses sub-word units and morpho-syntactic units for word segmentation. 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.

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.

We evaluate the "mixed" approach as a most appropriate for the 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.

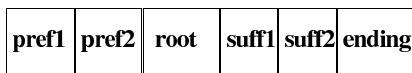


Figure 1: The word structure according to the general linguistic morphological theory.

2.2 Interpreting sound alternations

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.

The phonetic alternations in Bulgarian are of various types and influence both derivational and inflectional morphology. The general morphological theory offers a segmentation of words (Fig. 1) which consists of root to which prefixes, suffixes or endings are attached. In Bulgarian, all three types of morphemes are used and additional difficulties come from the fact that sound alternations can be occurred both in stems, prefixes, suffixes, and also on their boundaries which suggest extremely complicated solutions.

3 The Universal Networking Language

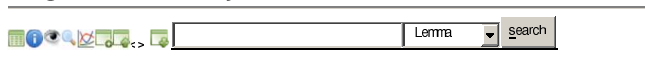
In the UNL approach, information conveyed by natural language is represented as a hypergraph composed of a set of directed binary labelled links (referred to as "relations") between nodes or hypernodes (the "Universal Words"(WS)), which stand for concepts (Uchida and Della Senta, 2005). UWs can also be annotated with "attributes" representing context information (UNL, 2011).

Universal Words (UWs) represent universal concepts and correspond to the nodes to be interlinked by "relations" or modified by "attributes" in a UNL graph. They can be associated to natural language open lexical categories (noun, verb, adjective and adverb). Additionally, UWs are organized in a hierarchy (the UNL Ontology), and are defined in the UNL Knowledge Base and exemplified in the UNL Example Base, which are the lexical databases for UNL. As language-independent semantic units, UWs are equivalent to the sets of synonyms of a given language, approaching the concept of "synset" used by the WordNet.

Attributes are arcs linking a node to itself. In opposition to relations, they correspond to one-place predicates, i.e., function that take a single argument. In UNL, attributes have been normally used to represent information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc). Attributes are annotations made to nodes or hypernodes of a UNL hypergraph. They denote the circumstances under which these nodes (or hypernodes) are used. Attributes may convey three different kinds of information: (i) The information on the role of the node in the UNL graph, (ii) The information conveyed by bound morphemes and closed classes, such as affixes (gender, number, tense, aspect, mood, voice, etc), determiners (articles and demonstratives), etc., (iii) The information on the (external) context of the utterance. Attributes represent information that cannot be conveyed by UWs and relations.

Relations, are labelled arcs connecting a node to another node in a UNL graph. They correspond to two-place semantic predicates holding between two UWs. In UNL, relations have

Bulgarian Dictionary



Class	UNL-NL Dictionary			NL Dictionary
	UWs	Lemmas	Polysemy	
Adjectives	288	509	1.77	388
Adverbs	171	229	1.34	199
Nouns	694	881	1.27	735
Verbs	529	843	1.59	682
Other	0	0	0.00	0
All	1682	2462	1.46	2004

Figure 2: The statistical word distribution of part-of-speech for UNL interpretation of Bulgarian inflectional morphology.

been normally used to represent semantic cases or thematic roles (such as agent, object, instrument, etc.) between UWs.

UNL-NL Grammars are sets of rules for translating UNL expressions into natural language (NL) sentences and vice-versa. They are normally unidirectional, i.e., the enconversion grammar (NL-to-UNL) or deconversion grammar (UNL-to-NL), even though they share the same basic syntax.

In the UNL Grammar there are two basic types of rules: (i) Transformation rules - used to generate natural language sentences out of UNL graphs and vice-versa and (ii) Disambiguation rules - used to improve the performance of transformation rules by constraining their applicability.

The UNL offers a universal language-independent and open-source platform for multilingual web-based applications (Boitet and Cardenosa, 2007) available for many languages (Martins, 2011) including Slavonic languages like Russian (Boguslavsky, 2005) as well.

3.1 Representing Bulgarian inflectional morphology in UNL

The UNL specifications offer types of grammar rules particularly designed to interpret inflectional morphology both with respect to prefixes, suffixes, infixes, and to sound alternations taking place during the process of the inflection. Thus, UNL allows two types of transformation inflectional rules: (i) A-rules (affixation rules) apply over isolated word forms (as to generate possible inflections) and (ii) L-rules (linear rules) apply over lists of word forms (as to provide transformations in the surface structure). Affixation rules are used for adding morphemes to a given base form, so to generate inflections or derivations. There are two types of A-rules: (i) simple A-rules involve a single action (such as prefixation, suffixation, infixation and replacement), and (ii) complex A-rules involve more than one action (such as circumfixation).

Bulgarian Grammar



Morphology

M14

езИК (ТИП 14) -К, -ЦИ 🗑️⚠️🔍🗨️

SNG&DEF:=0>"а"; SNG&DEF:=0>"ЪТ"; PLR:=1>"ци";
PLR&DEF:=1>"ците"; PAU:=0>"а"; MUL:=0>"а";

Figure 3: The inflectional rules definitions for the word "ezik".

There are four types of simple A-rules: (i) prefixation, for adding morphemes at the beginning of the base form, (ii) suffixation, for adding morphemes at the end of the base form, (iii) infixation, for adding morphemes to the middle of the base form, (iv) replacement, for changing the base form.

The analysed application of Bulgarian inflectional morphology (Noncheva and Stoykova, 2011) was made within the framework of the project 'The Little Prince Project' of the UNDL Foundation aimed to develop UNL grammar and lexical resources for several european languages based on the book 'The Little Prince'. Hence, the lexicon is limited to the text of the book. It offers the interpretation of inflectional morphology for the nouns, adjectives, numerals, pronouns (Stoykova, 2012) and verbs which uses A-rules (Fig. 2).

The UNL interpretation of nouns defines 74 word inflectional types. Every inflectional type uses its own rules to generate all possible inflected forms for the features of number and definiteness. Here we are analysing the inflectional rules of Bulgarian word for *language* "ezik"¹.

```
base form = ezik SNG&DEF:=0>"а";  
SNG&DEF:=0>"ut"; PLR:=1>"ci";  
PLR&DEF:=1>"cite"; PAU:=0>"а"; MUL:=0>"а";
```

The inflectional rules for generation of all inflected word forms are defined as separate rules (Fig. 3). The suffixation rules for adding: SNG&DEF:=0>"а";, SNG&DEF:=0>"ut";, PAU:=0>"а";, MUL:=0>"а"; use the idea of introducing stems to which the inflectional morphemes are added. A-rules for replacement also reflect the idea of introducing inflectional stems consisting of root plus infix PLR:=1>"ci";, PLR&DEF:=1>"cite";.

The generated inflected word forms of the example Bulgarian word for *language* "ezik" are given at the Fig.4. In general, the UNL lexical information presentation scheme underlie the idea of WordNet for semantic hierarchical representation and allows the presentation of synonyms and the translation of the word as well, which also is introduced in the application.

Adjectives are defined by using 14 word inflectional types and every inflectional type uses its own rules to generate all possible inflected forms for the features of gender, number and def-

¹ Here and elsewhere in the description we use Latin alphabet instead of Cyrillic. Because of mismatching between both some of Bulgarian phonological alternations are assigned by two letters instead of one in Cyrillic alphabet.

Bulgarian Grammar



Morphology

M14

език (тип 14) -к, -ци 🗑️⚠️🌐

base form=език SNG&DEF=езика SNG&DEF=езикът
PLR=езици PLR&DEF=езиците PAU=езика MUL=езика

Figure 4: The word forms of the word "език" generated by the system.

initeness. The interpretation of numerals and pronouns consist of 5 and 6 word inflectional types, respectively. Alternatively, verbs are represented in 48 inflectional types. The UNL interpretation, also, offers syntactic and semantic account. The syntactic account is represented by 21 syntactic rules for subcategorization frame and linearization, and rules to define the semantic relations.

In general, the UNL interpretation of Bulgarian inflectional morphology offers a sound alternations interpretation mostly by the use of A-rules. The inflectional rules are defined without the use of hierarchical inflectional representation even they define the related inflectional types. The sound alternations and the irregularity are interpreted within the definition of the main inflectional rule.

The UNL application, also, represents a web-based intelligent information and knowledge management system which allows different types of semantic search with respect to the context like semantic co-occurrence relations search, keywords or key concepts search, etc.

Conclusion

The demonstrated application of Bulgarian inflectional morphology uses the semantic networks formal representation schemes and the UNL as a formalism. However, it encodes the inflectional knowledge using both the expressive power and the limitations of the formalism used. The UNL knowledge representation scheme offers well defined types of inflectional rules and differentiates inflectional, semantic, and lexemic hierarchies. The treatment of inflectional classes as nodes in the inflectional hierarchy is used extensively, as well.

The application is open for further improvement and development by introducing additional grammar rules and by enlarging the database for the use in different projects.

References

(2011). URL <http://www.undl.org>.

Blevins, J. (2001). Morphological paradigms. *Transactions of the Philosophical society*, 99:207–210.

Boguslavsky, I. (2005). Some lexical issues of unl. In J. Cardenosa, A. Gelbukh, E. Tovar (eds.) *Universal Network Language: Advances in Theory and Applications. Research on Computing Science*, 12:101–108.

Boitet, C., B. I. and Cardenosa, J. (2007). An evaluation of unl usability for high quality multilingualization and projections for a future unl++ language. *In A. Gelbukh, (ed.) Proceedings of CICLing, Lecture Notes in Computer Sciences, 4394:361–376.*

Martins, R. (2011). Le petit prince in unl. *In Proceedings from Language Resources Evaluation Conference 2011*, pages 3201–3204.

Noncheva, V., S. Y. and Stoykova, V. (2011). The little prince project – encoding of bulgarian grammar. <http://www.undl.org>(UNDL Foundation).

Popov, K., G. E. and Penchev, J. (1983). *The Grammar of Contemporary Bulgarian Language (in Bulgarian)*. Bulgarian Academy of Sciences Publishing house.

Stoykova, V. (2012). The inflectional morphology of bulgarian possessive and reflexive-possessive pronouns in universal networking language. *In A. Karahoca and S. Kanbul (eds.) Procedia Technology*, 1:400–406.

Uchida, H., Z. M. and Della Senta, T. (2005). *Universal Networking Language*. UNDL Foundation.