Some Aspects of the Morphological Processing of Bulgarian

Milena Slavcheva
Linguistic Modelling Department
Central Laboratory for Parallel Processing
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St, 1113 Sofia, Bulgaria
milena@lml.bas.bg

Abstract

This paper demonstrates the modelling of morphological knowledge in Bulgarian and applications of the created data sets in an integrated framework for production and manipulation of language resources. The production scenario is exemplified by the Bulgarian verb as the morphologically richest and most problematic part-of-speech category. The definition of the set of morphosyntactic specifications for verbs in the lexicon is described. The application of the tagset in the automatic morphological analysis of text corpora is accounted for. A Type Model of Bulgarian verbs handling the attachment of short pronominal elements to verbs, is presented.

1 Introduction

The morphological processing of languages is indispensable for most applications in Human Language Technology. Usually, morphological models and their implementations are the primary building blocks in NLP systems.

The development of the computational morphology of a given language has two main stages. The first stage is the building of the morphological database itself. The second stage includes applications of the morphological database in different processing tasks. The interaction and mutual prediction between the two stages determines the linguistic and computational decision-making of each stage.

Bulgarian computational morphology has developed as the result of local (Paskaleva et al. 1993; Popov et al. 1998) and international activities for the compilation of sets of morphosyntactic distinctions and the construction of electronic lexicons (Dimitrova et al. 1998). The need for synchronization and standardization has lead to the activities of the application to Bulgarian of internationally acknowledged guidelines for morphosyntactic annotation (Slavcheva and Paskaleva 1997), and to the comparison of morphosyntactic tagsets (Slavcheva 1997).

In this paper I demonstrate the production scenario of modelling morphological knowledge in Bulgarian and applications of the created data sets in an integrated framework for production and manipulation of language resources, that is, the BulTreeBank framework (Simov et al. 2002). The production scenario is exemplified by the Bulgarian verb as the morphologically richest and most problematic part-of-speech category. The definition of the set of morphosyntactic specifications for verbs in the lexicon is described. The application of the tagset in the automatic morphological analysis of text corpora is accounted for. Special attention is drawn to the attachment of short pronominal elements to verbs. This is a phenomenon difficult to handle in language processing due to its intermediate position between morphology and syntax proper.

The paper is structured as follows. In section 2 the principles of building the latest version of

a Bulgarian tagset are pointed out and the subset of the tagset for verbs is exhaustively presented. Section 3 is dedicated to a specially worked out typology of Bulgarian verbs which is suitable for handling the problematic verb forms.

2 Morphosyntactic Specifications for Verbs As a Subset of a Bulgarian Tagset

2.1 Principles of Tagset Construction

The set of morphosyntactic specifications for verbs is a subset of the tagset (Simov, Slavcheva, Osenova 2002) used within the BulTreeBank framework (Simov et al. 2002) for morphological analysis of Bulgarian texts. A tagset for annotating real-world texts can be divided into several subsets according to the types of text units:

1. Tags attached to single word tokens. (These are the common words in the vocabulary: nouns, verbs, adjectives, adverbs, pronouns, prepositions, etc.)

2. Tags attached to multi-word tokens. (Most of them are conjunctions having analytical structure, but also some indefinite pronouns, etc.)

3. Tags attached to abbreviations.

4. Tags attached to named entities which are of various types: person names, topological entities, etc.; film titles, company names, etc.; formulas, single letters, citations of words as used in scientific texts.

5. Tags for punctuation.

The above groups of tags belong to three big divisions of tag types. The tags in items 1 and 2 above include annotation of linguistic items that belong to what is accepted to be a common dictionary. The tags in items 3 and 4 contain annotation of linguistic units that name, generally speaking, different kinds of realities and entities. They are peculiarities of the unrestricted, real-life texts. Item 5 refers to the annotation of linguistic items that serve as text formatters. The subject of discussion in this paper is a tagset of the first type, that is, morphosyntactic information attached to words as dictionary units.

As pointed above, the tagset for Bulgarian is constructed on the basis of the long term experience acquired in local and international initiatives for the compilation of core sets of morphosyntactic distinctions and the construction of electronic lexicons.

The EAGLES principle of levels of the morphosyntactic annotation is used but it is, so to say, localized. That means that while the EAGLES annotation schemes are constructed for the simultaneous application to many languages, in the tagset described here, the principle of levels is consistently applied for structuring the morphosyntactic information attached to each part of speech (POS) category in a single language, that is, Bulgarian. The principle of levels is applied in the EAGLES multi-lingual environment as follows. The elaboration of the tags starts with the encoding of the most general information which is applicable to a big range of languages (e.g., the languages spoken in the European Union). It continues with structuring the morphosyntactic information that is considered more or less common to a smaller group of languages. Finally, the single language-specific information is added to the tags. In the monolingual tagset for Bulgarian this scheme of information levels is used for the POS categories.

The next underlying structuring principle that is used in the Bulgarian tagset is that of MULTEXT and MULTEXT-East for ordering the information by defining numbered slots in the tags for the value of each grammatical feature and leaving the slot unoccupied if the feature is not relevant in a given tag. Again, in a multi-lingual environment, the MULTEXT ordering of the grammatical categories is simultaneously applied to a bunch of languages, while in the Bulgarian tagset it is applied to one and the same POS category of Bulgarian. The ordering of the information starts with the POS category and follows a scale of generality where the more general lexeme features (e.g. type, aspect of the verb) precede the grammatical features describing the wordform (e.g. person, number of verb forms).

The tagset is defined so that the necessary and enough information is attached to the word tokens

according to the following factors:

- The information is attached on the morphological level (that is, stemming from the lexicon).
- The information is attached to running words in the text (and here is the tricky interplay of form and function of the lexical items).
- There is potential for interfacing this information with the next levels of linguistic representation like, for instance, syntax (that is why we speak about morphosyntactic annotation).
- When defining the specifications in the tags, the levels of linguistic representation (i.e., morphology, syntax, semantics, and pragmatics) are kept distinct as much as possible. That means that the underlying principle is to provide information for the lexical items thinking about them as dictionary units. In connection to the latter principle, another principle is defined, that is, whenever possible, the formal morphological analyses of the lexical items are taken into account, rather than the assignment of functional categories to them, which is the task of the successive levels of linguistic interpretation and representation.

2.2 Format of the Tagset

The information encoded in the morphosyntactic annotation is represented as sets of feature-value pairs. The tags are lists of values of the grammatical features describing the wordforms. The format of the tags is a string of symbols (letters, digits or hyphens) where for each value there is one single symbol that denotes it. The first symbol is a capital letter denoting the POS category. The rest of the string is a mixture of small letters, digits or hyphens. The letters or digits denote the values of the features describing a lexical item. The hyphen means that a given feature is irrelevant for a given lexical item. The hyphen preserves the ordering of the values of features in the tag string by denoting a position. In case the hyphen or hyphens come last in the tag string, that is, no symbol follows them, they are omitted.

2.3 Specifications for the Verb

The grammatical features which are encoded in the verb tagset have ordered positions in the tag strings as shown bellow.

1:POS, 2:Verb type, 3:Aspect, 4:Transitivity, 5:Clitic attachment, 6:Verb form/Mood, 7:Voice, 8:Tense, 9:Person, 10:Number, 11:Gender, 12:Definiteness

All the descriptions below are in the form of triples where the first element is the name of the grammatical feature, the second element is the value of the grammatical feature and the third element is the abbreviation used in the tag string.

The feature-value pairs describing the verb category are distributed in three levels. The first level of feature-value pairs represents the most general category, that is, the part of speech.

[POS, verb, V]

The second level of description includes features whose values provide the invariant information for a given wordform, that is, the information stemming from the lexeme. This is the information used for the generation of the appropriate type and number of paradigm elements for a given lexeme. For the verb the second level features are: *Verb type, Aspect, Transitivity, Clitic attachment.* Combinations of those features denote subclasses of verbs. The features, their values, and the abbreviations are given in the following descriptions.

[Verb type, personal, p]
[Verb type, impersonal, n]
[Verb type, auxiliary, x]
[Verb type, semi-impersonal, s]
[Aspect, imperfective, i]
[Aspect, perfective, p]
[Aspect, dual, d]
[Transitivity, transitive, t]
[Transitivity, intransitive, i]
[Clitic attachment, none, 0]
[Clitic attachment, mandatory ”se”, 1]
[Clitic attachment, mandatory ”si”, 2]
[Clitic attachment, mandatory acc.pron., 3]
[Clitic attachment, mandatory dat.pron., 4]
[Clitic attachment, mandatory dat.pron.+se, 5]
[Clitic attachment, optional ”se”, 6]
[Clitic attachment, optional ”si”, 7]

The values of the third level features define the variant information for a given word form, that is,

the grammatical information carried by the various inflections. This is the level of most specific information. The third level features for the verb are: *Verb form/Mood, Voice, Tense, Person, Number, Gender, Definiteness.*

[Verb form/Mood, Finite_indicative, f]
[Verb form/Mood, Finite_imperative, z]
[Verb form/Mood, Finite_conditional, u]
[Verb form/Mood, Non-finite_participle, c]
[Verb form/Mood, Non-finite_gerund, g]
[Voice, active, a]
[Voice, passive, v]
[Tense, present, r]
[Tense, aorist, o]
[Tense, imperfect, m]
[Tense, past, t]
[Person, first, 1]
[Person, second, 2]
[Person, third, 3]
[Number, singular, s]
[Number, plural, p]
[Gender, masculine, m]
[Gender, feminine, f]
[Gender, neuter, n]
[Definiteness, indefinite, i]
[Definiteness, definite, d]
[Definiteness, Short_definite_form, h]
[Definiteness, Full_definite_form, f]

3 Type Model of Bulgarian Verbs and its Application in Lexicon Construction

The type model is the underlying factor in defining the morphosyntactic schemes for verbs and the scheme transformations necessary in different applications. Four initial Verb Types are defined: *personal, impersonal, semi-personal and auxiliary*. The definition of the types is triggered by the necessity to determine the relevant and optimal combinations of second level features which generate the correct paradigms of verbs belonging to the respective verb type. A decisive factor for the typology is the combination of verbs with short pronominal elements. It is necessary to differentiate, from one side, the attachment of short pronominals as an integral part of the lexeme for some groups of verbs (and consequently to the whole paradigm), and, on the other hand, the generation of combinations between verb forms and short pronominals when grammatical structures of various meanings come out.

At this point it should be noted that the electronic lexicon that is used for automatic morphosyntactic annotation in the BulTreeBank framework (Popov et al. 1998) follows the traditional, "paper-dictionary" subcategorization of verbs into personal, impersonal and auxiliary. Also the morphological analyzer identifies only single word tokens, that is, strings of symbols between white spaces. Orthographically, the short pronominal elements in Bulgarian are always separate word tokens and change their place around the verb according to language-specific phonological rules. In such a way, the full Type Model which takes into account the pronominal elements is, so to say, "switched off". It can be easily "switched on", since the full Type Model categories are subsets of the categories belonging to the model applied at present. In the BulTreebank tagset the slot for the values of the feature *Clitic_attachment* is filled by a hyphen, that is, the subcategorization according to the attachment of short pronominals is switched off, but it is easily recoverable when required.

Now let us consider the full Type Model proper and the templates of morphosyntactic specifications defined by the possible combinations of second level features, that is, features describing the invariant, lexeme information. The full Type Model is already applied in practice in the paradigmatic dimension of lexicon construction: 17909 Bulgarian verbs have been classified according to the model (Slavcheva 2002a).

3.1 Type Personal Verbs

The greatest number of verbs belong to this type. The personal verbs have a full paradigm of inflected forms. The number of the paradigm members depends on the features *Aspect* and *Transitivity*. The possible values of the feature *Clitic_attachment* are: *none, mandatory_se, mandatory_si*. In the working variant of the dictionary there exist the values *optional_se, optional_si* which are used for the generation of verb lexemes containing a reflexive formant (i.e., *se* or *si*).

The combination between a personal verb and the short accusative reflexive pronominal element

se defines the following classes of verbs:

1. Intransitive verbs with obligatory accusative reflexive element *se* (e.g., *usmihvam se* 'smile'), which have no correlates without *se*.

2. So called medium verbs (e.g., *karam se* 'quarrel') which have correlates without *se* (e.g., *karam* 'drive') but the meaning of two correlates is quite different. The short reflexive pronoun is not interchangeable with the full form of the reflexive pronoun *sebe si*.

3. Verbs denoting a reciprocal action (e.g., *bia se* 'fight').

4. Verbs that can be defined as reflexive *per se*, that is, the subject and the object of the action coincide. The subject is prevailingly animate. The interesting linguistic fact about those verbs is that, theoretically and logically, the alternation of short and full forms of the accusative reflexive pronoun is possible, but in reality the usage of the full form is communicatively very strongly marked and is not common at all. This fact supports the assumption that the combination between a verb and the short reflexive *se* is lexicalized (e.g., *aboniram* 'subscribe smb.' / *aboniram se* 'subscribe self').

The combination between a personal verb and the short dative reflexive pronominal element *si* defines the following classes of verbs:

1. Transitive and intransitive verbs with obligatory dative reflexive element *si* (e.g., *vaobraziavam si* 'imagine'), which have no correlates without *si*.

2. Medium verbs (e.g., *tragvam si* 'go home') which have correlates without *si* (e.g., *tragvam* 'go') but the meaning of two correlates is quite different.

3.2 Type Impersonal Verbs

The verbs belonging to this class have the smallest paradigm: the finite forms are only in the third person singular, and the participles are only in the neuter singular. The attribute *Transitivity* is irrelevant for the impersonal verbs. The possible values of the feature *Clitic_attachment* are: *none, mandatory_acc_pers_pron, mandatory_dat_pers_pron, mandatory_dat_pers_pron+se, mandatory_se*. The combination between an impersonal verb and the short pronominals results in the following classes:

1. Impersonal verbs without short pronominals (e.g., *samva* 'dawn').

2. Impersonal verbs with *se*, which are formal variants of some verbs belonging to class 1, that is, there is no difference in the meaning (e.g., *samva se* 'dawn').

3. Impersonal verbs with obligatory short accusative personal pronoun, short dative personal pronoun or short dative personal pronoun + se (e.g., *marzi me* 'to be lazy', *dozsaliava mi* 'to feel pitty', *gadi mi se* 'to feel sick'). The verbs in this class have no correlated forms of personal verbs without pronominals.

4. Impersonal verbs with short pronominals, which have correlated forms of personal verbs without short pronominals, but the attachment of the pronominals changes the meaning and triggers the differentiation of independent verb lexemes of impersonal verbs with pronominals (e.g, *trese* 'shake' / *trese me* 'be in a fever', *struva* 'cost' / *struva mi se* 'it appears to me').

3.3 Type Semi-personal Verbs

The definition of this innovative type of verbs is triggered by the idiosyncracies of the paradigm, the argument structure and the obligatory attachment of short personal pronouns. The verbs in this class have features in common both with the personal and the impersonal verbs and it is most convenient to isolate them in a separate class. The semi-personal verbs resemble the personal verbs in having a much bigger paradigm compared to the impersonal ones. In fact, forms in the first and second person singular and plural are not used (e.g, *vali* 'to rain', *boli me* 'it hurts me'). The semi-personal verbs can form sentences which structurally coincide with sentences of personal verbs, that is, they have a full-fledged subject, but the set

of nouns that can occupy the subject position is rather small, hence the argument structure is rather specific. (E.g., *Valiat porojni dazsdove.* 'Heavy rains fall.' *Krakata me boliat.* 'My legs hurt me.') The semi-personal verbs have also features in common with the impersonal verbs. They have the same possible combinations with the short pronominals as the impersonal verbs have. The feature *Transitivity* is irrelevant, as it is with the impersonal verbs. The subcategorization of the semi-personal verbs is analogous to that of the impersonal verbs (see items 1-4 for the impersonal verbs above).

3.4 Type Auxiliary Verbs

The small number of auxiliary verbs have an idiosyncratic paradigm. The features *Aspect*, *Transitivity* and *Clitic_attachment* are irrelevant for them.

4 Conclusion and Further Development

In section 3, the application of the verb Type Model in a paradigmatic dimension has been considered. A very important practical issue is how the morphosyntactic information encoded in the second level features (i.e., the lexeme information) can be used in a syntagmatic dimension, that is, pattern recognition and annotation in running texts. The issue of crucial importance is the utilization of the *Clitic_attachment* information.

Within the BulTreebank framework, a cascaded regular grammar has been built for the segmentation, pattern recognition and category assignment of Bulgarian compound verb forms as linguistic entities in XML documents (Slavcheva 2002b). In the segmentation model, the short pronominals are included into the compound verb forms of all types of verbs which consist of different combinations among short pronominals, particles and auxiliary verbs. At present the grammar for parsing compound verb forms does not discriminate between cliticized verb forms which are lexemes *per se* and cliticized verb forms which are purely grammatical. Thus an immediate application of the Type Model and the data set of approximately 18000 subcategorized verbs would be the construction of a discriminating parser for the different types of cliticized verb forms. In its turn, this more detailed morphosyntactic differentiation can be used as a source for predictions of the valency frame alternations in a machine-aided construction of the syntactic structure of sentences.

5 Acknowledgment

The work presented in this paper is supported by the BulTreebank project, funded by the Volkswagen Foundation, Federal Republic of Germany, under the Programme "Cooperation with Natural and Engineering Scientists in Central and Eastern Europe", contract I/76887.

References

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.J., Petkevič, V., Tufiş, D. (1998) "Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages." In *Proceedings of COLING-ACL'98*, Montréal, Québec, Canada, pp.315-319.

Paskaleva, E., Simov, K., Damova, M., Slavcheva, M. (1993) "The Long Journey from the Core to the Real Size of a Large LDB". In *Proceedings of ACL Workshop "Acquisition of Lexical Knowledge from Text"*, Columbus, Ohio, pp.161-169.

Popov, D., Simov, K., Vidinska, S. (1998) *Dictionary of Writing, Pronunciation and Punctuation of Bulgarian.* Atlantis LK, Sofia.

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001) "CLaRK - an XML-based System for Corpora Development." In *Proceedings of the Corpus Linguistics 2001 Conference*, pp.558-560.

Simov, K., P. Osenova, M. Slavcheva, S. Kolhovska, E. Balabanova, D. Doikov, K. Ivanova, A. Simov, M. Kouylekov. (2002) "Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank." In *Proceedings of LREC 2002*, Canary Islands, Spain, pp.1729-1736.

Simov, K., Slavcheva, M., Osenova, P. (2002) "BulTreeBank Morphosyntactic Tagset." BulTreeBank Report, Sofia, Bulgaria.

Slavcheva, M. (1997) "A Comparative Representation of Two Bulgarian Morphosyntactic Tagsets and the EAGLES Encoding Standard", TELRI I COPERNICUS Concerted Action 1202, Working Group 3 "Morphosyntactic Annotation", Report. http://www.lml.bas.bg/projects/BG-EUstand/

Slavcheva, M. (2002a) ”Language Technology and Bulgarian Language - Classificational Model of the Verb”. In *Proceedings of the Conference ”Slavistics in 21 Century. Traditions and Expectations.”*, SEMASH Publishing House, Sofia, Bulgaria, pp.240-247 (in Bulgarian).

Slavcheva, M. (2002b) ”Segmentation Layers in the Group of the Predicate: a Case Study of Bulgarian within the BulTreeBank Framework”. In *Proceedings of the International Workshop ”Treebanks and Linguistic Theories”*, Sozopol, Bulgaria, pp.199-209.

Slavcheva, M., Paskaleva, E. (1997) ”Application to Bulgarian. A contribution to the EAGLES Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages.” TELRI I COPERNICUS Concerted Action 1202, Working Group 3 ”Morphosyntactic Annotation”, Report. http://www.lml.bas.bg/projects/BG-EUstand/eagles/index.html