Integrating a Verb Lexicon into a Syntactic Treebank Production

Milena Slavcheva

Linguistic Modelling Department, Central Laboratory for Parallel Processing Bulgarian Academy of Sciences
E-mail: milena@lml.bas.bg

1 Introduction

The creation of linguistically interpreted corpora is a tedious task and the automation of the annotation process is indispensable. A fully automated annotation is hardly possible to achieve, since it requires very sophisticated and large knowledge bases which are, themselves difficult to create. However, a "machine-aided approach" to the annotating process, using as many as available sources of linguistic information is justified and desirable.

The creation of an HPSG-based syntactic treebank of Bulgarian [18] is a process of incremental augmentation of the real-world sentences with linguistic annotation as the result of linguistic analysis at different processing levels.

At this stage of the treebank production, there are the following main sources of linguistic knowledge that can be integrated into a grammar for automated chunk connection and assignment of grammatical relations:

1. the disambiguated fine-grained information included in the morphosyntactic tags attached to the word tokens [17], [21];
2. the segmentation of the sentences into chunks as the initial groupings of the lexical elements [20], [10], [11];
3. the verb subcategorization model [21] implemented in the lexical data base developed in Linguistic Modeling Department (LMD) [12], [13], [19].

This paper demonstrates to what extent the initial verb lexicon can enhance the automatic discrimination of some diathesis alternations of verbs. In this way, heuristics for the clause-level structures can be claimed which can foster the annotation tasks in building the syntactic treebank.
2 Methodology

An essential component of a parsing system are verb classes whose descriptions contain information about the relations of the predicates to their arguments. The importance of building data sets of verbs subcategorized according to the syntactic and semantic structures they form is proved by various projects [8], [4], [16].

Verbs can be grouped according to their syntactic behaviour, that is, the number and type of arguments they take in certain diathesis alternations. There is a general view that syntax is not enough in defining the principles and rules of the predicate - argument combinations. At the same time, it is the level of syntax that serves as the starting point for building large-scale semantic descriptors of verbs. Beth Levin, for instance, adheres to the position that meaning determines the syntactic behaviour of verbs, but in her well-known preliminary investigation of grouping the English verbs into semantic classes, she starts with the definition of diathesis alternations and the classification of verbs mainly in respect to those alternations [9]. Saint-Dizier, in his attempt to define verb semantic classes for French, suitable for implementation in NLP systems, relies on syntax which is more discernible in NLP applications compared to conceptual structures [16].

For English, there exist large-scale lexical resources which can be utilized in NLP applications: WordNet [5], Levin verb classes [9], FrameNet [1]. Although they are lexicographic in nature, the resources in question are a good basis for the development of computationally tractable resources, as can be seen in the VerbNet [8] and the PropBank [6], [7] projects.

For Bulgarian, there is no information so far for the achievements of the Balkanet project (i.e., WordNet of Balkan languages) [3]. There is not a monolithic study of the semantic and syntactic properties of the Bulgarian verbs like Levin classes of English verbs. There is a paper dictionary of the valency frames of Bulgarian verbs [15], but it consists of 1000 entries and needs extension and modification in order to be utilized in Language Engineering (LE) tasks. The recent transformation of the valency dictionary in a machine readable form [2] does not bring it much closer to an LE resource.

Therefore, the remaining possibility is to make as much use as possible of the available resources designed for LE applications.

For Bulgarian, a Type Model of the Bulgarian verbs is suggested [21] and applied in the LMD lexicon, which, therefore, contains a data set of 17909 verbs subcategorized by the key features Verb Type, Transitivity and Clitic Attachment [19]. The initial verb types are: personal, impersonal, semi-personal and auxiliary. The possible values of the Clitic Attachment feature are given in Table 1. The Clitic Attachment feature is included into the morphosyntactic specifications of the verb lexemes.
At this point it should be noted that the electronic lexicon used for automatic morphosyntactic annotation in the BulTreeBank framework [14] 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. In the BulTreebank tagset the slot for the values of the feature *Clitic Attachment* is filled by the `+` symbol, that is, the subcategorization according to the attachment of short pronominals is "switched off". It can be straightforwardly incorporated if required.

In fact, in all lexical databases for Bulgarian, the sets of morphosyntactic specifications encode information about lexical units that are single words. There have always arisen, though, the problem of representing units that structurally are multi-words, but lexically and grammatically are inseparable items. The nature of such linguistic units invokes the interplay between paradigmatic knowledge encoding within a lexicon, and real world realizations in sentences.

So, from one side, it is the construction of the lexicon that puts the question: "Which are the paradigmatically definable compound verb forms and what do we do with them within the lexicon?"

On the other side, it is the assignment of syntactic structures to sentences that evokes decision-making issues like:

- distinction of the levels of representation, for instance, lexical and phrasal;

- determination of the boundaries and components of meaningful syntagmatic patterns to be automatically recognized in real world texts, for instance, the combination between full-content verbs, short pronominals and auxiliary verbs;

- attachment of the patterns recognized at a lower level into syntactic units of a higher level, for instance, the attachment of NPs as complements to verbs.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clitic attachment</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>mandatory <em>se</em></td>
</tr>
<tr>
<td></td>
<td>mandatory <em>si</em></td>
</tr>
<tr>
<td></td>
<td>mandatory acc. pron</td>
</tr>
<tr>
<td></td>
<td>mandatory dat. pron</td>
</tr>
<tr>
<td></td>
<td>mandatory dat.pron+<em>se</em></td>
</tr>
<tr>
<td></td>
<td>optional <em>se</em></td>
</tr>
<tr>
<td></td>
<td>optional <em>si</em></td>
</tr>
</tbody>
</table>

Table 1: The *Clitic Attachment* feature
in order to build a VP;

- attachment of richer linguistic descriptions to the sentence structure, for instance, attachment of feature structure descriptions.

The goal of the present work is to explore and make use of the interaction between the lexicon and the sentence structure, manifested in encoding information about the predicate-argument structure in the lexicon and the specification of relations between the predicate and its arguments in sentence productions. The interplay between the lexicon and the diathesis alternations of verbs is examined specifically in relation to pronominal elements which are factors in the categorization of verb-centered structures.

3 Data

The combination between verb forms and short reflexive or non-reflexive pronominal elements in Bulgarian can be analyzed in terms of the interface between morphology and syntax, and the use of at least some degree of semantic interpretation as a first approximation to semantic representation proper. This is due to the remarkable content diversity embodied in the formal co-occurrence of a verb form and a short pronominal element. The short pronominal element is always a separate word, appearing either as a proclitic or enclitic depending on the linearization influence of the context and has different grammatical status regarding the type of syntactic structure that results from the interaction between a given verb form and the short pronominal element.

For instance, the reflexive accusative element se has diverse grammatical potential displayed on several linguistic levels. As a component of the verbal lexeme, it can produce a phonetic variant of a given intransitive or impersonal verb; or it can fulfill derivational morphological role in the formation of independent intransitive verbal lexemes with different semantics and argument selection compared to their "se-less counterparts". On the syntactic level the se element plays the role of a grammatical formative in passive, detransitive, anticausative, or impersonal constructions.

The interaction between a reflexive short pronominal and a verb with certain morphosyntactic specifications can have impact on different linguistic levels: morphological, lexical, syntactic, and even pragmatic.

For instance, the accusative reflexive se or the dative reflexive si can be an obligatory formal component of the lexeme of certain intransitive verbs (traditionally defined as verba reflexiva tantum).
The reflexive se or si attach to a great number of transitive verbs producing new lexemes with different semantics and different syntactic behaviour (middle verbs, in the most general interpretation of the term). The degree of semantic divergence between the pairs of se-verbs and se-less verbs, and between si-verbs and si-less verbs varies, but a generalization can be drawn that distinguishes the "reflexivised" verbs which are separate verbal lexemes.

The reflexive elements are employed in the generation of diathesis alternations like passive and impersonal.

The reflexive si very often expresses modality and thus is an element of the level of pragmatics.

### 4 Grammar

The combination between a full-content verb and a short pronominal element falls in two main categories:

1. **verb lexeme**, that is, a dictionary unit;
2. **verb form**, that is, grammatical unit realized in syntax, having specific function.

The grammar rules assign types of diathesis alternations predicted by an exhaustive combinatorics of feature values stemming from two sources of linguistic information:

1. **lexicon**: key values of features included in the morphosyntactic tags;
2. **syntagmatic patterns**: presence/absence of short pronouns within the boundaries of verb chunks, assigned by the application of a cascaded regular grammar.

The morphosyntactic specifications for verbs in the Bulgarian tagset used for morphological analysis are extensively discussed in Slavcheva [21]. In the present paper, the key values, used in the combinatorial grid are pointed out where necessary.

The classes of verb chunks recognized and marked in real-world texts are discussed in [20] and [22]. The relevant feature for the combinatorial grid is the presence of a short pronoun within the chunk of the full content verb. The absence of a short pronoun is meaningful and is indicated by the none value. The rest of the possible pronoun components of the verb chunks are named as follows: refl se, refl si, acc pron, dat pron, dat pron+se.
The types of diathesis alternations are inferred from the merging of the key specifications from the two sources of information: the lexicon and the chunks. The types of diathesis alternations are grossly defined and can be used as "first-pass predictions" of the relations between the predicate and its arguments. In the subsections that follow, let us consider the combinatorial grid and the predictions of diathesis alternations for several main classes of verbs.

4.1 Personal transitive verbs

In Table 2 the combinatorics for the personal transitive verbs is given. In the Lexicon column, all combinations of activated morphosyntactic specifications are provided. The possible values of the feature Clitic attachment, encoded in the lexicon are: none, mandatory se, mandatory si. The verb forms that are relevant are either finite, or active participles. The passive participle is considered separately, since it is morphologically marked for the passive. The Chunk column contains the specifications of pronominal ingredients, relevant for this specific combinatorial grid. The Prediction column includes the diathesis alternation, expected to be typically present in the text having in mind the merging of the Lexicon and Chunk specifications. For instance, the heuristics formulates the following rules for labelling the types of diathesis alternations.

1. In the morphosyntactic tag of a personal verb, if the value of the Clitic attachment feature is optional se and the value of the Transitivity feature is transitive, then the syntagmatic pattern verb+se introduces a middle construction, that is, the verb lexeme is, generally speaking, the middle counterpart of the transitive verbal lexeme (e.g., karam se 'quarrel' vs. karam 'drive').

2. In the morphosyntactic tag of a personal verb, if the value of the Clitic attachment feature is none and the value of the Transitivity feature is transitive, then the syntagmatic pattern verb+se introduces a passive construction.

3. The third row in Table 2 introduces an idiosyncratic phenomenon related to the reflexive si pronominal. It very often occurs as a verbal clitic, thus belonging to the verb chunk, but semantically expresses a kind of possession relation between the Agent in the Subject position and the Entity occupying the position of the Direct or Indirect Object. For example:

(1) Fond "13 veka Bulgarija" oshte [si tarsi] koraba.
Fund "13 centuries Bulgaria" still [refl look_for] ship-the.
'The "13 century Bulgaria" fund is still looking for its ship.'

The reflexive si as a verbal clitic also acts in many cases as a modal particle that adds specific nuances to the verb semantics.
Table 2: Combinatorial grid for personal transitive verbs

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Chunk</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>none, finite/active participle</td>
<td>none</td>
<td>active, transitive</td>
</tr>
<tr>
<td>none, finite/active participle</td>
<td>refl se</td>
<td>refl-passive</td>
</tr>
<tr>
<td>optional se, finite/active participle</td>
<td>refl si</td>
<td>modal/possessive</td>
</tr>
<tr>
<td>mandatory si, finite/active participle</td>
<td>refl se</td>
<td>middle</td>
</tr>
<tr>
<td>optional si, finite/active participle</td>
<td>refl si</td>
<td>active, transitive</td>
</tr>
<tr>
<td>optional si, finite/active participle</td>
<td>none</td>
<td>active, transitive</td>
</tr>
<tr>
<td>optional si, finite/active participle</td>
<td>refl se</td>
<td>active, transitive</td>
</tr>
</tbody>
</table>

4.2 Personal intransitive verbs

In Table 3 the combinatorics for the personal intransitive verbs is given. The combinations of lexicon specifications contain one more option in the sixth row of Table 3 which corresponds to verb lexemes containing the reflexive se as an obligatory element. Here are some of the rules for guessing the type of diathesis alternation for this specific lexical class of verbs.

1. In the morphosyntactic tag of a personal verb, if the value of the Clitic attachment feature is optional se and the value of the Transitivity feature is intransitive, then the syntagmatic pattern verb+se introduces a verb lexeme (rather than an impersonal construct of the counterpart intransitive lexeme without se, e.g., smeya ’dare’ vs. smeya se ’laugh’).

2. In the morphosyntactic tag of a personal verb, if the value of the Clitic attachment feature is none and the value of the Transitivity feature is intransitive, then the syntagmatic pattern verb+se introduces an impersonal grammatical construct (rather than a counterpart se-lexeme).

4.3 Impersonal verbs

The impersonal verbs, defined as such in the lexicon, can have values of the feature Clitic Attachment as follows: none, mandatory acc pron, mandatory dat pron, mandatory dat pron+se, mandatory se. Their identification in the text serves to filter out those predicates which are “immanently impersonal”, thus reducing the cases to be differentiated to personal verbs that are used in an impersonal alternation.

4.4 Semi-personal verbs

The definition of this innovative type of verbs is triggered by the idiosyncrasies of the paradigm, the argument structure and the obligatory attachment of short
personal pronouns. The verbs in this class have features in common with both 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 '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 values of the Clitic Attachment feature as the impersonal verbs have. The identification of the semi-personal verb lexemes has similar impact on guessing the diathesis alternations as the identification of the impersonal verb lexemes (see Section 4.3).

5 Experiment

As pointed above, the morphosyntactic tags are attached to single word tokens and characterize them as dictionary units. In the text, however, the combinations between verbs and short pronominals have features differing from the features of the single word ingredients.

An experiment has been performed to see to what extent the lexicon look-up of the value of the feature Clitic Attachment can serve as a discriminator of the types of diathesis alternations of verbs. In a test corpus of 9249 running words, the occurrences of verb chunks that include a personal transitive or personal intransitive verb and the short reflexive pronominal se have been matched to the list of verbs

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Chunk</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>none, finite/active participle</td>
<td>none</td>
<td>active, intransitive</td>
</tr>
<tr>
<td>none, finite/active participle</td>
<td>refl si</td>
<td>modal/possessive</td>
</tr>
<tr>
<td>optional se, finite/active participle</td>
<td>refl se</td>
<td>intransitive, se-lexeme</td>
</tr>
<tr>
<td>mandatory se, finite/active participle</td>
<td>refl se</td>
<td>intransitive, se-lexeme</td>
</tr>
<tr>
<td>mandatory si, finite/active participle</td>
<td>refl si</td>
<td>intransitive, si-lexeme</td>
</tr>
<tr>
<td>optional si, finite/active participle</td>
<td>none</td>
<td>intransitive, si-lexeme</td>
</tr>
<tr>
<td>optional si, finite/active participle</td>
<td>refl si</td>
<td>intransitive, si-lexeme</td>
</tr>
</tbody>
</table>

Table 3: Combinatorial grid for personal intransitive verbs
Table 4 provides the numbers relevant to the precision and recall measures:

- **all** is the total number of diathetic alternation assignments to the combinations *personal verb + se*;
- **cor** is the number of correct assignments;
- **miss** is the number of missed assignments due to missing lexical entries in the morphological analyzer.

For instance, in sentence (2), the chunk *se pribira* is correctly recognized as an occurrence of the intransitive verb lexeme *pribiram se* (‘to go home’), and not of the transitive verb lexeme *pribiram* (‘to gather’, ‘collect’).

(2) Toj [se pribira] sled vsichki nas.

He [refl go_home] after all us.

’He goes home after all of us.’

In example (3) the chunk *se puskat* is erroneously identified as an occurrence of the intransitive verb lexeme *puskam se* (‘to leave’): it is the passive form of the transitive verb lexeme *puskam* (‘allow’, ‘release’, ‘utter’).

(3) biletcheta ..., posle [se puskat] za vtoro utopreba.

tickets ..., afterwards [refl utter] for second use.

’tickets ..., after that are uttered for second use.’

The analysis of the errors instigates some revision of the specifications of the verb entries in the LMD lexicon, namely, the *Clitic attachment* values of some verbs have to be corrected. In the experiment, 10 of the incorrect assignments of diathetic alternation are due to an error in the lexicon.

### 6 Conclusion and further development

Although it is necessary to carry out experiments with a larger corpus, possibly containing texts of various genres (not only newspapers), the investigations performed so far prove the validity of the approach, that is, the heuristics for defining
diathesis alternations based on lexicon descriptors has its grounds and can be implemented in a guesser of diathesis alternation types. The output of the guesser partially predicts the syntactic relations of the predicate to its arguments expressed by NPs, PPs, etc., identified by the shallow processing module. A further development of the processes described in this paper is to build and evaluate an attacher of constituents based on the labelled types of diathesis alternations. Certainly, the refinement of the classes of diathesis alternations, as well as the definition of "shallow semantic" classes of Bulgarian verbs is necessary.

References


