Verb Valency Descriptors for a Syntactic Treebank

Milena Slavcheva

Bulgarian Academy of Sciences
Institute for Parallel Processing
25A, Acad. G. Bonchev St, 1113 Sofia, Bulgaria
milena@lml.bas.bg

Abstract

An essential component of Language Engineering (LE) tools are verb class descriptors that provide information about the relations of the predicates to their arguments. The production of computationally tractable language resources necessitates the assignment of types of predicate-argument relations to a great variety of verb-centered structures: it is necessary to define not only the initial, canonical valency frame of a great number of verb lexemes, but also the diathesis alternations, which reflect the real-life usage of verbs. This paper describes the implementation of descriptors of the valency properties of Bulgarian verbs used in the production of a syntactic treebank of Bulgarian. The descriptors are based on available LE resources for Bulgarian: a verb subcategorization model implemented in the lexical data base that is used; a chunk grammar that recognizes verb form patterns. Predictive models are built and applied in a grammar that annotates grammatical relations inferred from the combination of morphosyntactic and shallow syntactic processing cues. The real significance of this particular processing is the resolution, in relation to the valency properties of many verbs, of the discrepancy or the contradiction between the verb lexicon specifications and the verb syntagmatic realization.

1. Introduction

An essential component of Language Engineering (LE) tools are verb class descriptors that provide information about the relations of the predicates to their arguments (Kipper et al., 2000; Levin, 1993; Saint-Dizier, 1996; Walde and Brew, 2002). The production of computationally tractable language resources necessitates the assignment of types of predicate-argument relations to a great variety of verb-centered structures: it is necessary to define not only the initial, canonical valency frame of a great number of verb lexemes, but also the diathesis alternations, which reflect the real-life usage of verbs (Gildea and Palmer, 2002; Gildea, 2002; Lapata, 1999; Aranovich and Runner, 2000).

This paper describes the implementation of descriptors of the valency properties of Bulgarian verbs used in the production of a syntactic treebank of Bulgarian (Simov et al., 2002b). The descriptors are based on available LE resources for Bulgarian:

- a verb subcategorization model (Slavcheva, 2003c) implemented in the lexical data base that is used (Paskaleva et al., 1993; Paskaleva, 2003; Slavcheva, 2003a);
- a chunk grammar (Slavcheva, 2003b; Slavcheva, 2003d) that recognizes verb form patterns.

Predictive models are built and applied in a grammar that annotates grammatical relations inferred from the combination of morphosyntactic and shallow syntactic processing cues.

The grammar is regular and the rules are applied in a cascade (Simov et al., 2002a) where the levels of rule application are arranged in order of decreasing certainty with which the types of diathesis alternations can be identified. That is, successive filtering is performed where the initial indicators of the type of structures are key values of features included in the fine-grained lexicon specifications and the presence/absence of short pronouns within the verb chunks identified during the shallow parsing step.

The paper is structured as follows. Section 2 describes the grammar application. Section 3 accounts for the experiments carried out and provides an evaluation. Section 4 derives a conclusion and gives clues for further development.

2. Grammar

In the lexical database used, the sets of morphosyntactic specifications for verbs encode generalized lexical properties of single-word entries. For instance, in the lexicon, a verb is assigned the specification *transitive*, if it can take a direct object, even if that is possible only in some of its usages. A verb is *intransitive*, if it does not take a direct object: nothing is said about the prepositional phrases and/or other constituents that it combines with. A specific phenomenon is the combination of Bulgarian verbs with clitic pronouns: the problem arises of representing units that structurally are multi-words, but lexically and grammatically are inseparable items. Thus quite formal indicators are at play in the space of application of the current grammar. Formal indicators like those, however, individually, or in combination, induce syntactically and grammatically motivated classifications. The hypothesis of the connection between the meaning of a verb and its syntactic behaviour has been extensively explored, asserted and utilized (Levin, 1993; Saint-Dizier, 1996; Walde and Brew, 2002).

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 key features in the lexicon are *Verb Type*, *Transitivity*, *Clitic Attachment*, *Verb Form/Mood* and *Voice*. The latter two features are necessary for selecting the proper verb forms for the grammar rules: the finite verb forms and the active form of the participles. Table 1 provides only the values of the key features that are relevant for the grammar.

Feature	Value
Verb Type	personal impersonal semi-personal
Transitivity	transitive intransitive
Clitic attachment	none mandatory se mandatory si mandatory acc. mandatory dat. mandatory dat.+se optional se optional si
Verb Form/Mood	Finite_indicative Finite_imperative Non-finite_participle
Voice	active

Table 1: Key feature-value pairs from the lexicon

In the verb chunks, the relevant feature for the grammar is *presence/absence* of a specific short pronoun. The values of that feature are given in Table 2.

Value	Description
none	clitic pronoun absent
se	acc. reflexive clitic pronoun present
si	dat. reflexive clitic pronoun present
acc	acc. personal clitic pronoun present
dat	dat. personal clitic pronoun present
dat+se	combination dat. pers. pron & se

Table 2: Key values from the chunks

Predictions of the diathesis alternations (DA) are inferred from the exhaustive combinatorics of the key features from the lexicon and the key features from the chunks. It should be noted that the types of diathesis alternations are grossly defined and used as ”first pass predictions” of the relations between the predicate and its arguments. It should be emphasized that the real significance of this particular diathesis alternation assignment is that the grammar output resolves the initial valency properties of the verb as a lexical unit realized syntagmatically in a sentence production (Slavcheva, 2003b).

The syntagmatic co-occurrence of verbs and clitic pronouns results in the formation of verb complex structures whose initial diathesis type can be predicted with different certainty depending on the Verb Type. For instance, the labelling of the ”inherently impersonal” verb chunks filters out the certain cases of impersonal diathesis. The cases of personal verbs used in an impersonal alternation can be recognized with a smaller degree of certainty using the mechanisms in the current grammar.

Table 3 represents the types of diathesis alternations, assigned by the current grammar rules. The columns entitled Verb Type, Transitivity and Clitic Attachment contain information stemming from the lexicon. The Chunk column contains information stemming from the verb chunk. The DA column includes the predictions of the diathesis alternation types. In Table 3, the diathesis alternations are arranged in order of decreasing certainty of the predictions. At present, the certainty of the diathesis alternation assignment is intuitive (for instance, it is not formalized as probability of error (Manning and Schuetze, 1999)) and roughly groups the diathesis alternations in the following manner:

- Diathesis alternations from number 1 to number 17 in Table 3 are assigned with high certainty.
- Diathesis alternations from number 18 to number 23 are assigned with a lower degree of certainty compared to those of 1-17.
- Diathesis alternations from number 24 to number 32 are of the lowest degree of certainty compared to the previous two groups.

3. Experiment

An experiment has been performed to assess the grammar that discriminates initial types of diathesis alternations as defined in the current predictive model. In a test corpus (newspaper texts) of 14830 running words, the current grammar has been applied. The heuristics for labelling the types of diathesis alternations are provided in Table 3. The evaluation measures given in Table 4 are calculated for the occurrences of *personal* verb chunks that contain a reflexive pronominal *se*, or a reflexive pronominal *si*, that is, the output of the guesser for diathesis alternations of type 15-17 and 24-32 (see Table 3) is examined.

The numbers relevant to the precision and recall measures are given in Table 4. The *true positives (tp)* are the cases of correctly assigned types of diathesis alternations. The *false positives (fp)* are the cases of wrong assignments. The *false negatives (fn)* are cases that failed to be selected by the grammar due to lapses in the morphosyntactic annotation.

tp	fp	fn	Precision	Recall
257	52	6	83.17%	97.71%

Table 4: Evaluation of diathesis discrimination

No	Verb Type	Transitivity	Clitic Attachment	Chunk	DA
1	impersonal	irrelevant	none	none	impersonal
2	impersonal	irrelevant	optional se	none	impersonal
3	impersonal	irrelevant	optional se	se	impersonal
4	impersonal	irrelevant	mandatory se	se	impersonal
5	impersonal	irrelevant	mandatory acc	acc	experiential
6	impersonal	irrelevant	mandatory dat	dat	experiential
7	impersonal	irrelevant	mandatory dat+se	dat+se	experiential
8	semi-personal	irrelevant	none	none	semi-personal
9	semi-personal	irrelevant	optional se	none	semi-personal
10	semi-personal	irrelevant	optional se	se	semi-personal
11	semi-personal	irrelevant	mandatory se	se	semi-personal
12	semi-personal	irrelevant	mandatory acc	acc	experiential
13	semi-personal	irrelevant	mandatory dat	dat	experiential
14	semi-personal	irrelevant	mandatory dat+se	dat+se	experiential
15	personal	intransitive	mandatory se	se	intransitive
16	personal	intransitive	mandatory si	si	intransitive
17	personal	transitive	mandatory si	si	transitive
18	personal	transitive	none	none	transitive
19	personal	transitive	optional se	none	transitive
20	personal	intransitive	none	none	intransitive
21	personal	intransitive	optional se	none	intransitive
22	personal	intransitive	optional si	none	intransitive
23	personal	transitive	optional si	none	transitive
24	personal	intransitive	optional se	se	middle
25	personal	intransitive	none	se	impersonal
26	personal	transitive	none	se	passive
27	personal	transitive	optional se	se	middle
28	personal	intransitive	optional si	si	middle
29	personal	transitive	optional si	si	middle
30	personal	intransitive	none	si	modal/possessive
31	personal	transitive	none	si	transitive & modal/possessive/si-dative
32	personal	transitive	optional se	si	transitive & si-dative

Table 3: Diathesis alternation predictions

The analysis of the errors provides useful feedback related to the specifications in the lexicon: the values of the *Clitic attachment* feature of some verbs are reassessed and they will be possibly corrected. Within the test corpus, the number of *Clitic attachment* specifications evaluated as incorrect is 18. In case they are corrected, the output of the guesser will differ in favour of precision.

Factors that influence the operation of the grammar are:

- low frequency of occurrence of the diathesis alternations that have the highest degree of certainty (i.e., diathesis alternations numbered 1-17 in Table 3);
- the generalization of the types of diathesis alternations, which is especially apparent in diathesis alternations from number 26 to number 32 in Table 3, or these are the diathesis alternations with the lowest degree of certainty.

It should be noted that the current model does not include the combination between a personal verb and the clitic accusative and dative pronouns. This stems from the fact, that this particular combination is not considered in the lexicon due to its syntactic regularity which renders it as improper for a lexicon of the format used in the present work.

4. Conclusion and further development

The investigations carried out so far prove the validity of the approach, that is, the predictive models for diathesis alternation assignment based on lexicon descriptors have their grounds and are utilizable in a guesser of initial diathesis alternation types. The real significance of this particular process is the resolution, in relation to the valency properties of many verbs, of the discrepancy or, in many cases, the contradiction between the verb lexicon specifications and the verb syntagmatic realization. The next processes are related to the refinement of the verb classes in the lexicon, the definition of more specific types of diathesis alternations, and a ”shallow semantic” subclassification of Bulgarian verbs.

The utilization of entirely supervised methodology is motivated by the characteristic features of the available LE production for Bulgarian. The advantages of this particular methodology lie in its prognostic power, that is, the modelling and the resource production at a given level is projected to the successive levels.

5. References

Aranovich, R. and Runner J T., 2000. Diathesis Alternations and Rule Interaction in the Lexicon. *WCCFL20 Proceedings*, ed. K.Megerdoomian and L.A. Barrel, Somerville, MA: Cascadilla Press, pp.101-114.

Gildea, D., 2002. Probabilistic Models of Verb-Argument Structure. *Proceedings of COLING'02*

Gildea, D. and Palmer M., 2002. The Necessity of Parsing for Predicate Argument Recognition. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp.239-246.

Kipper, K., Dang, H. T., and Palmer, M., 2000. Class-Based Construction of a Verb Lexicon. *Proceedings of AAAI-2000: Seventh National Conference on Artificial Intelligence*, Austin, Texas.

Lapata, M., 1999. Acquiring Lexical Generalizations from Corpora: A Case Study for Diathesis Alternations. *Proceedings of the 37th Annual Meeting of ACL*, Maryland, USA, pp.397-404.

Levin, B., 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press, Chicago and London.

Manning, C. D., Schuetze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts and London, England.

Paskaleva, E., 2003. Automatic Processing of Bulgarian and Russian Language Resorces in Consecuitve and Parallel Mode. *Proceedings of the Conference "Slavistics at the Beginning of 21 Century. Traditions and Expectations."*, SemaRSH Publishing House, Sofia, pp.111-120.

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

Saint-Dizier, P., 1996. Constructing Verb Semantic Classes for French: Methods and Evaluation. *Proceedings of COLING'96*, Denmark, pp.1127-1130.

Simov, K., Kouylekov, M., Simov, A., 2002a. Cascaded Regular Grammars over XML Documents. *Proceedings of the Second Workshop on NLP and XML (NLPXML-2002)*, Taipei, Taiwan.

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

Slavcheva, M., 2003a. Language Technology and Bulgarian Language - Classificational Model of the Verb. *Proceedings of the Conference "Slavistics at the Beginning of 21 Century. Traditions and Expectations"*., SemaRSH Publishing House, Sofia, pp.209-216.

Slavcheva, M., 2003b. Corpus Shallow Parsing: Meeting Point between Paradigmatic Knowledge Encoding and Syntagmatic Pattern Matching. *Proceedings of "Corpus Linguistics 2003" Conference*, Lancaster, UK.

Slavcheva, M., 2003c. Some Aspects of the Morphological Processing of Bulgarian. *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*, Budapest, Hungary, pp.71-77.

Slavcheva, M., 2003d. Extracting Verb Complex Structures in Bulgarian. In Cunningham, H., Paskaleva, E., Bontcheva, K., Angelova, G. (eds.), *Proceedings of the International Workshop on Information Extraction for Slavonic and Other Central and EasternEuropean Languages, RANLP 2003*, Borovets, Bulgaria, pp.94-101.

im Walde, S. S. and Brew C., 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp.223-230.