Corpus shallow parsing: meeting point between paradigmatic knowledge encoding and syntagmatic pattern matching

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

1. Introduction

The production of language corpora augmented with linguistic information takes place in an environment where the researcher is constantly faced by the problem of the interaction between the paradigmatic knowledge representation and the syntagmatic properties of real-life language usage. This problem is especially apparent in the interaction between the first two steps in the development of a syntactic treebank, that is, the result of the morphosyntactic annotation of texts and the shallow parsing aiming at identifying reliable and meaningful segments in the text. Usually, when one comes across the problem of the interdependence between language as a system (i.e., langue) and language as real-life means of communication (i.e., parole), the attention is drawn to entities specific for the text like dates, titles, names, addresses (Karttunen et al. 1997), which are absent in a conventional lexicon used for morphosyntactic analysis. In general, more research has been dedicated to named entities and nominal objects than to the predicate. Similar observation is expressed in Nenadic et al. 2001. Certainly, named entities are those anchors in the text which, on one hand, have to be processed in order to resolve the problem of full text annotation, and, on the other hand, constitute, so to say, the substantive content of the text. The same holds for the noun groups in general. In this paper, however, I concentrate on the predicate of natural language sentences which is, so to say, the event information centre in texts. Since verbs are in the backbone of the lexicon and the grammar, it might seem at first glance, that they smoothly "shuttle" between the paradigmatic and syntagmatic dimensions of language formal description. But this is not the case. My target of exploration are Bulgarian verbs and the two aspects of difficulty that they represent: 1) complex forms expressing tense, mood and voice, which involve auxiliary verbs; 2) verbal units including clitic reflexive and personal pronominals, as well as different types of particles. The analysis and description of those entities are necessary for the appropriate assignment of the syntactic structure of sentences. They serve as the basis for the incremental augmentation of annotated texts with information about the temporal and discourse positioning of events, as well as of the argument structure of predicates.

The work presented in this paper is a component of the integrated BulTreeBank framework (Simov et al. 2002b) where grammars are built for the segmentation, pattern recognition and category assignment of linguistic entities in XML documents. The software environment is provided by the CLARK system (Simov et al. 2001) which incorporates a number of tools for the creation and manipulation of XML documents, a cascaded regular grammar engine and a system of constraints over XML documents.

The construction of a grammar for the segmentation, pattern recognition and category assignment of Bulgarian compound verb forms is carried out in an iterative, incremental mode, that is, the source and the target of grammar rule compilation and application are interchanged so that the grammar is refined and its discriminating power is increased.

The advantages of the approach described in this paper lie in the exploitation of well-established and relatively simple techniques, that is, regular expressions and finite-state automata in a multi-task unified framework for the creation and manipulation of linguistically interpreted data sets. In this way, with reasonable efforts, satisfying results are obtained so that the results of given processes are considered relatively complete for use in certain applications, and, at the same time, there is the constant possibility for linguistically meaningful augmentation of the data sets within an underlying infrastructure.

The paper is structured as follows. In Section 2 the relevant data categories are overviewed. Section 3 provides a survey of the sources of linguistic knowledge and the computational devices necessary in the initial phase of grammar writing. Section 4 presents the rule sets for automatic recognition and category

assignment of compound verb forms. Section 5 provides some preliminary evaluation of the grammar application. In Section 6 some conclusions are drawn which serve as the basis for the further development of the grammar. Section 7 summarises some points in the further grammar development.

2. An overview of the data categories

The first problem that has to be solved in building a grammar for automatic recognition of compound verb forms is to determine the boundaries and the components of the linguistic entities that stand for the patterns to be recognized. The factors that determine the decision-making are the following:

- language-specific idiosyncrasies;
- shallow parsing strategies as a module in the text corpus processing;
- interface between the segments identified as a result of shallow parsing and the deeper linguistic analysis performed at subsequent stages of the creation of the treebank.

2.1. The tense, mood and voice paradigm of Bulgarian verbs

Verbs in Bulgarian have a very rich tense, mood and voice paradigm consisting of simplex (synthetic) inflected forms and complex (analytic) forms. The latter are generally composed of a non-finite form of the full-content verb and one or more (in some cases omitted) auxiliaries. The reader can get an impression of the number and variety of forms if I mention that, traditionally, Bulgarian is considered to have nine tenses and four moods, including the "exotic" narrated (not-witness) mood. Passive voice is represented by the form consisting of auxiliary verbs and the passive participle of full-content verbs. (The forms with the reflexive pronominal *se* are not included in the passive voice paradigm in the current model). In fact, no matter how modern theoretical linguists might determine the different number and types of tenses, moods and voices in Bulgarian, the forms are there to challenge the computational linguist in grammar writing.

Examples 1, 2 and 3 bellow illustrate some complex verb forms:

1) Toj shte e napisal pismoto.
   He will-aux,pres,3p,sg be-aux,pres,3p,sg write-active_aorist_particip,m,sg letter-the
   'He will have written the letter.'

2) Toj bil napisal pismoto.
   He be-past_participle,m,sg write-active_aorist_participle,m,sg letter-the
   'They say, he has written the letter.'

3) Knigite sa bili namereni pod masata.
   Books-the be-aux,pres,3p,pl be-past_participle,pl find-passive_particip,pl under table-the
   'The books have been found under the table.'

The difficulties in processing the complex tense, mood and voice forms stem from the fact that they incorporate features of both morphology and syntax. The morphological features are manifested in the grammatical meaning carried by the entire unit consisting of auxiliaries and a full-content verb. The syntactic features lie in the multi-word structure of the grammatical unit: the word order of the separate tokens in the construct can be permutated and different "external" syntactic elements can be inserted in-between the ingredients of the complex verb form.

2.2. The relation between verbs and small words

In Bulgarian, short pronominal elements and particles (which will be called *small words* for convenience) surrounding the verbs turn out to be a specific problem to encoding linguistic information in the lexicon, to sentence segmentation at the shallow parsing level, as well as to the phrase structure descriptions at the level of deeper linguistic analysis. In the segmentation model presented in this paper, the small words

accompanying the verbs belong to the verb form patterns. The small words are the following: the negative particle *ne*, the interrogative particle *li*, the preverbal element *da*, the accusative reflexive pronominal element *se*, the dative reflexive pronominal element *si*, the short accusative and dative personal pronouns.

The arguments for the inclusion of small words into compound verb forms can be outlined as follows. Both in the paradigmatic and syntagmatic dimension, the negative and interrogative particles, the preverbal element *da*, and the reflexive pronominals produce positional and meaningful clusters with the verb. The same holds for the combination between some impersonal verbs and accusative or dative pronominals expressing the number and person of the experiencer, thus producing a morphological entity with the verb. The arguments for the inclusion of short accusative and dative pronominals into compound forms with personal verbs are sought at the interface of shallow processing and deeper linguistic analysis. Object doubling is a phenomenon typical for Bulgarian. This is the case when the direct or indirect object is expressed twice: once by the short accusative or dative pronominal, and once by a full-fledged nominal phrase, which can be a full form of a pronoun, or a noun phrase. Oversimplifying the analysis, it is claimed in this paper that the short pronominals and the full-fledged complements are attached at different levels: the short pronominals at the lexical level, and the full-fledged complements at the syntactic level. In Avgustinova 1997, a similar approach to the segmentation of the Bulgarian verb complex is described. Similar analysis (although at a deeper linguistic level) of pronominal clitic attachment on a presyntactic level is observed for Italian (Monachesi 1998) and for French (Miller and Sag 1997). Personal pronominal clitics are part of the verb chunk in the shallow parsing grammar of French described in Clement and Kinyon 2000.

3. Sources of linguistic knowledge and facilities for grammar writing

Let us see what we have at the moment of being assigned the task of constructing a grammar for parsing compound verb forms. The team of the BulTreeBank project provides a corpus of one million word tokens which has the following special-purpose characteristics. The texts are from newspapers and are organised in a set of XML documents supplied with TEI conformant markup of the text structure at the paragraph level. The texts are processed by a morphological analyser and are manually disambiguated using the facilities of the system of constraints incorporated in ClaRK (Simov et al. 2002a). The electronic lexicon (Popov et al. 1998) used for the morphosyntactic analysis contains lexical entries for single words only. Consequently, the information about the verb tense, mood and voice is reduced to those portions, which are contained in single verb forms. In such a way three verb tenses (*present, aorist and imperfect*) are encoded in the tags attached to full-content verbs. Mood is represented by the imperative forms of full-content verbs and some special conditional forms of the auxiliary verb *sam* ('to be'). Voice is present as a category in part of the paradigm of personal transitive verbs, that is, the passive participle ending in *–n(a,o,i)*. There is no information about the attachment of short personal and reflexive pronominals, neither negative nor interrogative forms.

The purpose of the grammar is the automatic identification of simplex or complex verb forms, which constitute the predicate of the sentence. Throughout this paper I will use the term *compound verb form* to refer to those verb clusters which consist of more than one orthographic word, that is, the compound verb forms are combinations of a full-content verb and a different number and type of auxiliary verbs, short pronominals or particles in varying word order.

The grammar is determined by the ClaRK system environment (Simov et al. 2002a): it is of type regular, accepted by finite-state automata, and the rules are applied in a cascade where the output of a certain set of rules is the input to another set of rules (except for the initial rules). The rules of the grammar are of the type

C -> R,

where R is a regular expression and C is a category of the pattern matched by R.

The process of grammar construction starts with abstracting over grammar books, which, as it is well-known, are written for humans and do not partition or combine the linguistic knowledge in a way that can directly serve practical applications. Grammar books and paper dictionaries do not contain the suitable exhaustive sets of data structures used in real-life software applications. Nevertheless, using grammar books and one's competence of language, a grammar is constructed in a deductive mode of inference. The result of the application of this grammar can be defined as a first approximation to resolving the problem of parsing. The grammar writer has to define principles of grammar construction to lead him in the "oceans of

language" (Sampson 2003), and in this particular case in the ocean of Bulgarian verb forms. The principles of the initial phase of grammar construction can be defined as follows.

1. *Exhaustiveness.* The grammar writer thinks of all constructs that can count as forms of tense, mood and voice: simplex forms, various occurrences of combined finite and non-finite auxiliary and main verb forms. Naturally, the positive forms are considered first, but then it is necessary to enumerate the forms of predicate negation and interrogation expressed by the addition of the negative particle *ne* and the interrogative particle *li* respectively. The variation of constructs grows rather big when the short pronominals are added into the set of ingredients of compound forms.

2. *Definition of key morphosyntactic features and combinations of features*. The morphosyntactic tags attached to the verb tokens in the text are combinations of values of a whole bundle of morphosyntactic features differentiating the members of the verb paradigms. In the input words to the regular expressions, it is necessary to specify those values of features that are relevant and necessary for the program to perform the correct pattern matching. For instance, the lexical features *Aspect* and *Transitivity* are irrelevant, and they are denoted by wildcards in the input words to the regular expressions. At the same time, it is necessary to mention explicitly the indefinite forms of the past active and of the passive participles since they can be components of tense, mood or voice forms. Thus the definite forms which can be parts of noun or adjective phrases are discarded as possible ingredients of verb clusters. (The morphological expression of definiteness in Bulgarian is a definite article in the form of an ending to word forms.) The full description of the morphosyntactic specifications for verbs as a subset of the BulTreeBank tagset (Simov et al. 2002c) is given in Slavcheva 2003.

3. *Continuity.* It is a linguistic fact that the continuity of compound verb forms can be interrupted at certain points and different elements, mainly adverbials and nominals can be inserted in them. But in the initial phase of grammar development, compound verb forms are thought of as prevailingly continuous entities, rather than discontinuous ones. This is due to the mainly paradigmatic perspective of thinking at the initial phase of modeling the verb complex. The alternative of foreseeing the eventual occurrence of "external" elements in-between the ingredients of the compound verb forms using regular expressions in an unsupervised manner does not help much (Nenadic 2001; Zackova and Pala 1999). It is necessary to work out carefully the set of rules that envisage the discontinuity of compound verb forms. That is possible after a closer and more extensive examination of syntagmatic patterns (Zackova and Pala 1999; Nenadic 2001).

4. *Longest versus shortest match principle*. The cascaded regular grammar engine of the ClaRK system provides the possibility to identify segments of different length in a sequence suitable for defining the structure and supplying the annotation of compound verb forms. A strategy is defined for choosing the most suitable length and composition of the examined constructs at intermediate levels of parsing.

4. Rules and category assignment at the initial phase of grammar development

The methodology described in section 3 is illustrated by some rules for syntagmatic pattern matching.

The first instance is a rule that recognises segments consisting of a full-content verb and a certain number and type of small words.

```
<MV>\w</MV> -->

<"da","#">?, <"ne","#">?, <"#","Pp@d@@@t">?, <"li","#">?,
<"#","Pp@a@@@t">?, <"li","#">?, <"#",("Ppxa@@@t"|"Ppxd@@@t")>?,
<"#",("Vp@@@f#"|"Vp@@@z#"| "Vn@@@f#"|"Vp@@@cao@@@i"|
"Vp@@@cam@@@i"|"Vp@@@cv@@@@i"| "Vn@@@cao@@@i"| "Vn@@@cam@@@i")>,
<"li","#">?,<"#","Pp@d@@@t">?, <"#","Pp@a@@@t">?,
<"#",("Ppxa@@@t"|"Ppxd@@@t")>?
```

The right hand side of the rule is a regular expression stating that a full-content verb cluster can consist of:

1) a finite verb form inflected for the present, aorist or imperfect tense;
2) a non-finite verb form which can be an indefinite form of the active aorist, active imperfect, or the passive participle;
3) one of the verb forms enumerated in items 1 and 2 preceded or followed by a variety of sequences of small words.

The left hand side of the rule contains the XML markup that is added to the output of the rule, that is, the tags that enclose the recognized pattern denoted by the variable \w. Since the input words to the regular expression of the rule are the contents of XML elements nested into XML local trees, an Element Value in the form of an XPath expression (Xpath 1999) is defined for the element serving as the context node of the tree structure that is computed by the grammar. In the case of the MV rule described here, the element (marked with <w>) is the context node for which the following XPath expressions are written:

ph/text() ta/text()

The XPath expressions state that the local tree structure should be searched for the PCDATA strings of the *ph* element (whose content is the running word string itself), and for the PCDATA strings of the *ta* element (whose content is the correct morphosyntactic tag) in order to match the input words to the regular expression. In the rule itself, the input words to the regular expressions are ordered pairs of *ph* content and *ta* content. The irrelevant content for a given input word is designated by the # symbol, for instance, <"da","#">, or <"#","Pp@d@@@@t">. The grammar also indicates the elements in the XML document to which the grammar is applied. In the current example the MV rule is applied to the paragraph (p), head (head) and hilighted (hi) elements. This is indicated by the expression

//p|//head|//hi

At present the regular grammar cascade for recognition of compound verb forms consists of two levels. The set of rules that constitute Recognizer 1 and operate on the first level includes rules that produce:

- verb forms consisting of a single main verb or a main verb accompanied by small words;
- verb forms consisting of a single auxiliary verb or auxiliary verb accompanied by small words;
- groups of small words, enclosed in a separate chunk, which remain unattached to a verb due to special cases of linear order and discontinuity.

The second-level set of rules (Recognizer 2) outputs segments corresponding to full tense, mood and voice forms and including small words. For instance, the following rules belong to Recognizer 2:

<VC>\w</VC> -> <CT>?, <XV>+, <MV>

<VC>\w</VC> -> <MV>, <XV>+

The input words to the regular expressions on the right hand side of the rules are tags denoting XML elements that are added at Level 1 of pattern recognition. MV stands for a main verb chunk, XV for an auxiliary verb chunk, and CT for an independent chunk of small words. The regular expressions encode the variants of verb complex composition. It should be noted that the rules overgenerate verb complex constructs since they lack a detailed description of the possible verb complex constructs in relation with subordinate clauses. The left hand side of the rules state that the category of the patterns recognized by the rules are of type verb complex and the markup that delimits it in the XML document consists of a tag named VC.

5. Some preliminary evaluation of the grammar application

The application of Recognizer 2 provides the possibility to delimit correctly and to mark up the majority of the longest compound verb form patterns. The experiments with the newspaper corpus show that the chunks identified by Recognizer 1 tend to be adjacent within the longest compound verb form patterns in communicatively unmarked written prose. A paradigmatic representation of word order within the verb complex in view of the "communicative organization of Bulgarian sentences" provided in Avgustinova 1997 supports the conclusion expressed in this paper.

Here are some numbers that illustrate the application of the regular grammars described in this paper. In a text of 4292 running words, there are:

- 536 occurrences of main verbs with or without small words recognized by the set of rules of Recognizer 1;
- 164 occurrences of auxiliary verbs with or without small words recognized by the set of rules of Recognizer 1;
- 5 occurrences of small word groups recognized by Recognizer 1 as a separate chunk due to specific linear order in respect to the main verb.

The application of Recognizer 2 resulted in the recognition of 77 compound verb forms *per se* which are longest matches. It should be noted that in the current grammar the combination between the copula and the primary predicatives is not identified as a pattern to be recognized by the grammar. They remain separate entities. The 77 occurrences of verb complexes were manually evaluated and the result was that in one case the verb complex was erroneously identified as such: the verbal part of a subordinate clause was incorrectly combined into a verb complex with a preceding copula. In the cases of discontinuous compound verb forms the grammar simply fails.

The results of the grammar application to a number of text documents shows that the ratio of occurrences among the different types of segments is preserved.

6. Some conclusions

The first phase of the grammar development for parsing compound verb forms provides valuable input to the subsequent phases. Several conclusions are drawn which serve as the basis for the further development of the grammar.

1. The complex tense forms consisting of auxiliaries and a full-content verb are not of such a great number and variety in real-life texts and that is why the grammar has good performance. At the same time, it is necessary to augment the paradigmatic knowledge by learning from the rare occurences of syntagmatic patterns, since this is the only way to build a wide-coverage grammar of compound verb forms.

2. Discontinuous compound verb forms with adverbial and nominal inserts are not recognised by the grammar. They can be searched for due to the application of rules that identify shorter segments within the verb complex: auxiliary verb chunks and main verb chunks. In such a way the syntagmatic realisation compound verb forms serves as the source for paradigmatic knowledge necessary for the grammar rule writing. The direction from the syntagmatic to the paradigmatic dimension in grammar construction is also followed in the lemmatisation of compound verb forms in Serbo-Croatian in a method using local grammars described in Nenadic 2001. It is also obvious that discontinuity has to be explored in relation to the category of the main verb form. For instance, the passive participles are separated by adverbials from the accompanying auxiliary verbs in a greater number of cases compared to past active participles.

3. In the process of building the treebank, a core set of sentences, excerpted from Bulgarian grammar books, are manually assigned syntactic structure. This is a resource for the account of possible compound verb forms which include or not small words, and whose components are adjacent or

discontinuous. In such a way the possible inserts in compound verb forms can be classified and the grammar can be augmented with rules for the recognition of discontinuous compound verb forms as well (Zackova and Pala 1999).

7. Further development

The processes of grammar development within an integrated XML framework described in this paper provide a flexible method for augmenting data sets with linguistic information that is required as necessary for a certain application or research. The grammar developer can decide upon the degree of specification or underspecification of categories, the size and type of linguistic entities that should be analyzed. The decision making at the level of shallow parsing is performed in view of possible interfacing with the level of sophisticated linguistic analysis which is a subsequent step in building the treebank of Bulgarian. Having defined the underlying principles of grammar construction for the recognition of compound verb forms, it is necessary to perform more intensive testing of the application of grammar rules and to provide a detailed account of the frequency of occurrence of the different types of constructs. The exploration has to be performed in interrelation with the grammars developed by the team members for the other linguistic entities like noun groups, adjective and adverb phrases, named entities, etc. This is the way to extract the idiosyncratic cases where the grammar fails and to refine the grammar so that the deviating structures are possibly processed automatically as well. A significant further development is the interfacing between the shallow parsing level and the sophisticated linguistic analysis so that the process of treebank construction is facilitated.

Acknowledgement

The work presented in this paper is carried out within 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

Avgustinova T 1997 *Word order and clitics in Bulgarian*. Saarbruecken Dissertations, University of the Saarland.

Clement L, Kinyon A 2000 Chunking, marking and searching a morphosyntactically annotated corpus for French." In *Proceedings of ACIDCA'2000*, Monastir.

Karttunen L, Chanod J-P, Grefenstette G, Schiller A 1997 Regular expressions for language engineering. In *Natural Language Engineering*, Cambridge University Press, pp.1-24.

Miller P, Sag I 1997 French clitic movement without clitics or movement." *Natural Language and Linguistic Theory:15*, pp.573-639.

Monachesi P 1998 Decomposing Italian clitics. In Malari, S and Dini, L (eds) *Romance in HPSG*. CSLI Publications, Stanford, USA, pp.305-357.

Nenadic G, Vitas D, Krstev C 2001 Local grammars and compound verb lemmatization in Serbo-Croatian. In Zybatow G, Junghanns U, Mehlhorn L, Szucsich L (eds) *Current issues in formal Slavic linguistics*. Peter Lang: Europaischer Verlag der Wissenschaften, pp 469-477.

Popov D, Simov K, Vidinska S. 1998 *Dictionary of writing, pronunciation and punctuation of Bulgarian*, Atlantis LK, Sofia.

Sampson G 2003 Thoughts on two decades of drawing trees. In Abeille A (ed), *Treebanks. Building and using parsed corpora*. Kluwer Academic Publishers, Boston/Dordrecht/London, pp 23-41.

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

Simov K, Osenova P, Slavcheva M, Kolhovska S, Balabanova E, Doikov D, Ivanova K. Simov, A. Kouylekov M 2002b 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 2002c *BulTreeBank morphosyntactic tagset*. BulTreeBank Report, Sofia, Bulgaria.

Slavcheva M 2003 Some aspects of the morphological processing of Bulgarian. In *Proceedings of EACL Workshop on morphological processing of Slavic languages*, Budapest (in press).

XPath 1999. XML Path Language (XPath), Version 1.0. W3C Recommendation. http://www.w3.org/TR/xslt

Zackova E, Pala K 1999 Corpus-based rules for Czech verb discontinuous constituents. In *Proceedings of TSD'99 Workshop*, LNAI 1692, Springer-Verlag, pp.325-328.