Segmentation Layers in the Group of the Predicate: a Case Study of Bulgarian within the BulTreeBank Framework*

Milena Slavcheva
Linguistic Modelling Department, Central Laboratory for Parallel Processing
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev St, 1113 Sofia, Bulgaria
Phone: (+359 2) 979 28 12, Fax: (+359 2) 70 72 73
milena@lml.bas.bg, http://www.BulTreeBank.org

Abstract

This paper describes the development of a regular grammar that automatically recognizes and delimits segments in the group of the predicate in sentences of Bulgarian. The language-specific segmentation is performed at the level of partial parsing where reliable, meaningful and useful entities are formed called chunks. The significance of the grammar development lies in the fact that it is a plug-in into the BulTreeBank framework for the creation and exploitation of linguistically interpreted data sets. The results of the work presented in this paper are the initial stage in the development of a model for the description and processing of the group of the predicate at subsequent linguistic information levels.

1 Introduction

The main goal of the work presented in this paper is to develop and implement a regular grammar that automatically recognizes and delimits segments in the group of the predicate in sentences of Bulgarian. The language-specific segmentation is performed at the level of partial parsing where reliable, meaningful and useful entities are formed called chunks. The significance of the grammar development described in this paper lies in the fact that it is a plug-in into an entire framework for the creation and exploitation of linguistically interpreted data sets. Thus the task fulfilled is to study the properties of the Bulgarian verbs and their dependants and to provide the adequate representation at different levels of linguistic analysis ranging from chunking to sophisticated linguistic descriptions. Each representation level requires adequate processing and addition of annotation in an XML-based architecture implemented in a specific software environment. In this way the linguistic representation at each level is worked out in view of its interplay with the other levels and the overall technological infrastructure.

The paper is structured as follows. Section 2 outlines the settings of the framework that conduct the grammar development. Section 3 presents the argumentation for the linguistic decisions that determine the segmentation of the group of the predicate in Bulgarian. Section 4 describes the regular grammar set for the Bulgarian verb complex as part of the BulTreeBank grammar architecture.

*This work is funded by the Volkswagen Stiftung, Federal Republic of Germany, under the Programme "Cooperation with Natural and Engineering Scientists in Central and Eastern Europe" contract I/76 887.

2 Research and Development Settings

The grammar development described in this paper depends on the settings of the framework within which it is carried out, that is, the BulTreeBank project [12]. The modelling of the predicate and its group is determined by the structure of the BulTreeBank language data, the methodology of adding the linguistic interpretation and the software environment used in the creation and exploitation processes.

The BulTreeBank language data consist of three interrelated sets. The first one is a core set of 2500 manually analyzed sentences used as the golden standard for the linguistic annotation. The second set is the treebank *per se* consisting of syntactically annotated sentences excerpted from a text corpus. The foreseen size of the treebank is one million running words. The third set is a text corpus of 100 million words consisting of XML documents which contain multi-layered linguistic annotation, added incrementally to the text data. The first layer is paragraph level TEI [15] conformant encoding of the logical structure of texts. The second layer is morphosyntactic annotation attached to the word tokens. The subsequent layers of encoding of linguistic information are gradually achieved by the development and application of partial grammars, which recognize and assign markup to different meaningful segments in the texts. It is this particular level of processing that the grammar described in this paper is an integral part of.

The methodology of building the linguistically interpreted data is defined by the interplay of several underlying assumptions. A rule-based approach is generally applied in the data processing at the different layers ranging from the partial parsing to the most sophisticated level of representation, which is the attachment of HPSG conformant feature structure descriptions. The software environment is XML-based and is supported by the CLaRK system [13], 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.

3 The Language Data and the Segmentation Principles

The chunking of sentences is based on the philosophy of easy-first parsing outlined by Abney in [2, 3] The parsing technique uses reliable patterns consisting of categories and regular expressions that enter finite-state automata operating in the so called cascade, that is, sequence of levels of phrase recognition. The usefulness of partial, but robust, rapid and relatively accurate parsers has been recognized in large-scale text processing applications [7, 11, 8].

The notion of chunk is formulated in such a way that the grammar writer is provided with the necessary freedom to decide which those robust, reliable, meaningful and useful segments are in the sentences of his/her own language [2]. The search for chunks at a certain step of language processing raises the problem of the applicability of the chunk philosophy. In [1], Abney, discussing the properties of chunks, states that:

”There are constraints on the formation and interdependency of chunks. They are the Chunk Connectedness and the Chunk Inclusiveness constraints.”

And then he continues:

”Chunk Connectedness is not a universal constraint; it does not hold in German, for instance. Alternatively, we might speculate that German lacks chunks, rather than Chunk Connectedness. Conceivably, chunks fill a gap left by an impoverished morphology, providing polymorphemic word-like units in language with scant inflection or agglutination. However that may turn out, the crosslinguistic characterization of chunks is an important issue for further research.”

This statement dates back to 1990. Since then the chunk ideology has been applied to different languages (e.g., German [11], French [6]).

A careful observation shows that the factors that define chunks are, in fact, taken from different linguistic levels. The factors cannot be defined as purely prosodic, although prosody has its impact on the positioning of words and thus on chunk formation. They are not purely morphosyntactic or syntactic, although constituency and headship influence the structure of chunks. They are not pure word order factors, although adjacency and discontinuity are in most cases decisive in defining chunk boundaries. The factors cannot be defined as purely functional, although the grouping of words into chunks is often decided in view of the mutual functional interdependence among them.

In this paper I concentrate on one specific step in the VP modelling, that is, the definition of the grammar rules for chunk parsing of the verb complex as a component of the group of the predicate. In fact, the verb phrase (VP) as a maximal projection of the verb (V), consisting of verb forms and complement phrases, remains uncompleted at the chunk level. At this particular level of processing, it is necessary to define:

1. What components should the VP constructs be stripped off?
2. How should the remaining components be grouped into chunks, that is, reliable, non-recursive, meaningful and useful segments?

The answer to the first question above is the definition of the target of exploration and formal description in this paper, that is, constructs usually named in the literature as verb complexes. In Bulgarian the categories that build the verb complexes are:

- main verbs;
- auxiliary verbs and functional words that participate in the formation of complex analytic tense, mood and voice forms;
- elements generally referred to as clitics: short forms of accusative and dative personal pronouns, short reflexive pronouns, the negative particle, the interrogative particle and the preverbal element *da*.

At the level of chunking, the full-fledged complements of the verbs are not included into the verb chunks. They constitute their own chunks that are connected to the verb complexes at the level of deeper syntactic analysis. Chunk stylebooks developed for other languages document a similar approach [4, 11, 6].

The answer to the second question above provides the chunk model of the verb complex *per se*. The verb complex construct in Bulgarian consists of two main zones:

- main verb zone;
- auxiliary verb zone.

Verbs in Bulgarian have a very rich tense and mood paradigm consisting of simplex (synthetic) and complex (analytic) verb forms. The simplex (one-word) verb forms are inflected forms of content verbs or inflected forms of auxiliary verbs when used alone and functioning as full verbs. For instance, the present, aorist and imperfect tense forms are simplex ones.

Examples:

(1) Penka [napisa] pismo.
Penka [write-aorist,3p,sg] letter
'Penka wrote a letter.'

(2) Toj [e] dobar prijatel.
He [be-pres,3p,sg] good friend
'He is a good friend.'

Complex (multi-word) tense, mood and voice forms are a combination of finite or non-finite forms of the main verbs and one or more finite or non-finite forms of auxiliary verbs that constitute the auxiliary verb zone. The auxiliary zone may contain some forms of impersonal verbs and functional words.

Examples:

(3) Decata [sa] veche [napisali] sachinenieto.
Children-the [be-pres,3p,pl] already [write-participle,active,aorist] essay-the
'The children have already written the essay.'

(4) [Shtjaha] dosega [da sa] [doshli].
[Shta-aux,imperfect,3p,pl] until now [da-particle be-pres,3p,pl] [come-part,act,aor]
'They would have come until now.'

In Bulgarian, clitics surrounding the verbs turn out to be a specific problem to sentence segmentation at the partial parsing level, as well as to the phrase structure descriptions at the level of deeper linguistic analysis. The segmentation model presented in this paper is built on the basis of several underlying linguistic decisions. The verbal clitics are part of the main verb chunk, or, in some cases, some clitics are part of the auxiliary chunk.

Let us consider the arguments for such a decision for the different types of clitics. First of all, it should be pointed out that the term clitics is used here in its more general sense, that is, it refers to one-syllable words that attach to other words in the sentence to form prosodic entities but the emphasis is not on the phonological properties of clitics examined in detail as part of the phonological level, but rather on the function of clitics as part of the morphosyntactic and syntactic level of description. Having in mind the nature of clitics, prosody determines their position and ordering within the sentence and in this way influences the decisions taken about the chunking of sentences, but in the present model the prosodic properties of clitics are taken for granted and are neither the only, nor the decisive factor for the decision making.

The first goal of chunking is to identify those segments in the sentence that are "islands of stability", that is, clusters of words that, when present in the sentence, usually appear in a fixed position in relation to one another, forming some kind of an entity.

The negative particle *ne* always precedes the verb form it attaches to and only pronominal clitics can be inserted between the particle and the verb form. In case of simplex tense forms the negative particle is adjacent to the main verb, while in complex forms it is adjacent to the auxiliary verb. Thus the attachment of the negative particle is defined by the following rule: *ne* belongs to the chunk of the verb that it immediately precedes or precedes in a distance occupied by one or more pronominal clitics.

Examples:

(5) Uchiteljat [ne dojde] vchera.
Teacher-the [not-particle come-aorist,3p,sg] yesterday

’The teacher did not come yesterday.’

(6) Uchiteljat [ne beshe] oshte [doshal].
Teacher-the [not-particle be-aux,past,3p,sg] yet [come-participle,active,aorist]
’The teacher had not come yet.’

The interrogative particle *li* immediately follows the verb it attaches to in case there are no other clitic elements. The interrogative particle can be adjacent to the main verb or to the auxiliary verb in some complex forms. When there are clitic elements surrounding the verb, *li* is idiosyncratically ordered in respect to the other clitics. The interrogative particle belongs either to the main verb chunk or to the auxiliary verb chunk. For example:

(7) [Dojde li] uchiteljat vchera?
[Come-aorist,3p,sg not-particle] teacher-the yesterday
’Did the teacher come yesterday?’

(8) [Beshe li] uchiteljat [doshal] veche?
[Be-aux,past,3p,sg li-particle] teacher-the [come-participle,active,aorist] already
’Had the teacher come already?’

(9) [Ne mu go li dade] vchera?
[Not-pc him-clitic,masc,3p,sg it-clitic,3p,sg li-particle give-aorist] yesterday?
’Didn’t you give it to him yesterday?’

(10) [Ne mu li go dade] vchera?
[Not-pc him-clitic,masc,3p,sg li-particle, it-clitic,3p,sg give-aorist] yesterday?
’Didn’t you give it to him yesterday?’

The preverbal element *da*, as its name suggests, always precedes the verb and can be separated from it only by pronominal clitics. Depending on the tense and mood expressed by the verbal form in the sentence, *da* is attached either to the main verb or to the auxiliary verb, and belongs to the respective chunk.

Examples:

(11) Uchiteljat [uspja] vchera [da dojde].
Teacher-the [manage-aorist,3p,sg] yesterday [da-particle come-aorist,3p,sg]
’The teacher managed to come yesterday.’

(12) Uchiteljat [shteshe] veche [da e]
Teacher-the [shta-aux,imperfect,3p,sg] already [da-particle be-aux,pres,3p,sg]
[doshal].
[come-participle,active,aorist]
’The teacher would have come already.’

The accusative reflexive pronominal element *se* either precedes or follows the main verb depending on the immediate context in which the verb occurs. As a rule it is adjacent to the verb, except in cases of complex verb forms in the third person when the auxiliary verb *sam* is involved. The reflexive element *se* belongs to the chunk of the main verb, except the cases when it precedes or follows the

auxiliary verb: then it forms its own pron chunk. The analysis of the variety of syntactic alternations and semantic connotations triggered by the attachment of *se* to the verbs is beyond the scope of this paper, but in order to show the range of initial constructs that are captured by the grammar described in this paper, it is worth mentioning that the *se* forms of the verbs, besides being medial and reflexive verbs, can express impersonality and passivisation.

(13) Uchitelite [sa] [se vurnali] veche.
Teachers-the [be-aux,pres,3p,pl] [se-refl return-participle,active,aorist,pl] already
'The teachers have returned already.'

(14) Uchiteljat [se] [e] [vurnal] veche.
Teacher-the [se-refl] [be-aux,pres,3p,sg] [return-participle,active,aorist,sg] already
'The teachers have returned already.'

(15) [Vurnal se] [e] uchiteljat veche.
[Return-participle,active,aorist,sg se-refl] [be-aux,pres,3p,sg] teacher-the already
'The teacher has returned already.'

(16) [Vurnali] [sa] [se] uchitelite veche.
[Return-participle,active,aorist,pl] [be-aux,pres,3p,pl] [se-refl] teachers-the already
'The teacher has returned already.'

The position of the dative reflexive pronominal element *si* in respect to the verb is similar to that of *se*. Besides the formation of some medial and reflexive verbs, *si* functions mainly as a modal particle assigning special attitude of the speaker to the action expressed by the verb. Being a modal particle, *si* can occur in cases when the sentence has only an auxiliary verb in the role of a full verb.

Examples:

(17) [Otivam si] naj-posle.
[Go-pres,1p,sg si-refl] at last.
'I am going home at last.'

(18) Ivan [si] [e] tuk.
Ivan [si-refl] [be-aux,pres,3p,sg] here.
'Ivan is here.' (In the sense that Ivan is at home. "Home" is understood in a general abstract sense: "at his own place", or "in his homeland".)

The short accusative and dative personal pronouns require special consideration. The decision of putting them in one chunk with the main verbs is supported by several arguments.

The first question that arises is why, at all, the accusative and dative pronominal clitics are grouped in a chunk together with the verb. The answer includes the following arguments:

- There are impersonal verbs to which the accusative or dative clitics, or the combination dative clitic + *se* are obligatorily attached to express the person and number of the experiencer, thus forming a morphosyntactic entity with the verb.
- Arguments 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 accusative or dative clitic, and once by a full-fledged nominal phrase, which can be a full form of a pronoun, or a noun phrase. Oversimplifying the analysis, I will just mention that the pronominal clitics and the full-fledged complements are attached at different levels: the clitics at the word level, and the full-fledged complements at the syntactic level. Tanya Avgustinova, although with different goals, describes a similar approach to the segmentation of the Bulgarian verb complex in [5]. She distinguishes types of verb clusters within the verb complex that involve pronominal clitics. Similar analysis (although at a deeper lingustic level) of clitic attachment on a presyntactic level is observed for Italian [10] and for French [9]. Personal pronominal clitics are part of the verb chunk in the shallow grammar component of the French text corpus described in [6].

The second question that arises is why pronominal clitics belong to the chunk of the main verb and not to the chunk of the auxiliary verb. The answer is supported by the following arguments:

- The pronominal clitics in Bulgarian are usually adjacent to the main verb even if auxiliary verb forms are present. In this respect Bulgarian differs from Italian and French where clitic climbing is observed in the presence of auxiliary verbs.
- Having in mind the interface with the more sophisticated level of linguistic description, pronominal clitics satisfy the subcategorization requirements of the main verbs and they are the primary component of meaningful clusters formed by the main verbs. An exception is the case when the sentence contains an auxiliary verb only which is in the role of a full verb, and there is a dative clitic which functions as a modal particle revealing specific attitude to the event expressed by the auxiliary verb, or is a preverbal clitic having the meaning of possession. In this case the pronominal clitic forms its own pron chunk.

Examples:

(19) Penka [mu go dade] uchebnika na Ivan.
Penka [him-clitic,dat,masc,3p,sg it-clitic,acc,masc,3p,sg give-aorist] textbook-the to Ivan
'Penka gave the textbook to Ivan.'

(20) [Shteshe] Penka [da mu go
[Shta-aux,imperfect,3p,sg] Penka [da-particle him-cl,dat,m,3p,sg it-cl,acc,m,3p,sg
dade].
give-aorist]
'Penka would have given it to him.'

(21) Toj [mi] [e] prijatel.
He [to-me:clitic,poss,1p,sg] [be-aux,pres,3p,sg] friend.
'He is a friend of mine.'

The general principles that determine the formation of verb complex chunks at the lowest segmentation layer can be informally formulated as follows:

1. The main verb chunk and the auxiliary verb chunk can consist of one element (i.e., a simplex form of a main verb or an auxiliary verb), or more than one element.

2. In case there are more than one auxiliary verb forms in the sentence, each one of them belongs to a separate auxiliary chunk.

3. If the pronominal clitics (one or more than one) are immediately adjacent to the main verb (either preceding or following it), they belong to the chunk of the main verb.

4. If the pronominal clitics (one or more than one) are separated from the main verb (either preceding or following it) by one or more elements (which can be auxiliary verbs, functional words, particles, or phrases), they form their own chunk, the pron chunk.

5. The negative particle *ne* and the preverbal element *da* belong to the chunk of the verb form (either main or auxiliary) that they precede. In the various cases of ordering and discontinuity between the clitic forms and the verb forms, they follow the attachment principles characteristic of pronominal clitics.

6. The interrogative particle *li* belongs to the chunk of the verb form that it follows.

4 Regular Grammar Set for the Bulgarian Verb Complex

The construction of the regular grammar that recognizes verb complex patterns is conditioned by the CLaRK software environment which provides a cascaded regular grammar processor [14]. The partial grammars that delimit linguistically meaningful entities are applied on XML documents, that is, specific techniques are necessary for the formulation of the input words to the grammar and the incorporation of the output into the hierarchy of XML elements. The input words to the regular expressions of the grammar are the contents of XML elements. An implementation of the XPath language [16] is used to denote the relevant content nested in a local XML tree. A system of built-in and user-defined tokenizers are also provided by CLaRK in order to distinguish tokens in the XML documents that are suitable for the grammar applications.

The construction of the verb complex grammar is also conditioned by the processes of linguistic analysis and annotation that precede the chunking step. The text contains morphosyntactic information in the form of tags attached to each word token as a result of the application of an automatic morphological analyzer and manual disambiguation facilitated by the system of constraints incorporated in CLaRK [14]. Thus the input words to the input level to the regular grammar cascade, Level 0, consist of:

- the content of the *true analysis (ta)* element which is of type PCDATA and is a disambiguated morphosyntactic tag;

- the content of the *phonology (ph)* element which is of type PCDATA and is a naturally occurring word token.

In the grammar rules for verb complex chunks, the morphosyntactic tags used as components of the input words to the regular expressions are those attached to the verbs and to the short pronominal forms. For instance, *Vpit+f-r3s* is a string of the abbreviated values in the following list of attribute-value pairs:

[Category : Verb, Verb_type : personal, Aspect : imperfective, Transitivity : transitive, Clitic_ attachment : indicated, Verb_form / Mood : finite_indicative, Voice : irrelevant, Tense : present, Person : third, Number : singular].

Special wildcard symbols are used to underspecify those values in the morphosyntactic tags, which are irrelevant for the pattern matching operation. For example, if we want the grammar to recognize as a sub-word the present tense form of a given main verb, we write the above mentioned tag as *Vp@@@f#*, where the wildcard symbol @ stands for exactly one symbol and # stands for zero or more symbols. Thus this regular expression corresponds to the values in the following list of attribute-value pairs:

[Category : Verb, Verb_type : personal, Aspect : any, Transitivity : any, Clitic_attachment : any, Verb_form / Mood : finite_indicative, Voice : irrelevant, Tense : present, Person : any, Number : any].

The content of the *phonology (ph)* element is used in the verb complex grammar as a component of the input word to the regular expressions when the category referred to is a functional one and its morphosyntactic tag, indicating only the part of speech, cannot be used to specify the necessary element. For instance, all particles are tagged with *T* (abbreviating "Particle") and in order to denote the negative particle *ne* so that it is recognized as part of the verb chunks, it is necessary to refer to the concrete, naturally occurring word.

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.

Let us consider a rule that recognizes main verb chunks.

```
<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 the chunk of the main verb 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 (the definite article in Bulgarian is in the form of an inflection) 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 including short forms of accusative and dative personal pronouns, short forms of the accusative and dative reflexive pronouns, the preverbal element *da*, the negative particle *ne*, the interrogative particle *li*.

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 documents nested into XML local trees, an Element Value in the form of an XPath expression 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 expression is written:

```
ph[text(4,n,("Da"|"da"|"Ne"|"ne"|"Li"|"li"))]/text()|
ta[../ph[not(text(4,n,("Da"|"da"|"Ne"|"ne"|"Li"|"li")))]]/text()
```

The XPath expression states that the local tree structure should be searched for the PCDATA strings *da*, *ne*, or *li* which are the content of the *ph* element, and for the PCDATA strings which are the content of the *ta* element for the words in the regular expression other than the particles *da*, *ne* and *li*, or these are the morphosyntactic tags of the verbs and the pronominal clitics. 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
```

The CLaRK grammar editing tool also provides the possibility to specify a left and a right Regular Expression that denote the context in which the target regular expression occurs.

The set of rules whose input words belong to Level 0, and whose output are chunks at Level 1, constitute Recognizer 1. At the present stage of grammar development Recognizer 1 includes rules that produce:

- main verb chunks enclosed in <MV> and </MV> tags;
- auxiliary verb chunks enclosed in <XV> and </XV> tags;
- pron chunks enclosed in <CL> and </CL> tags.

Here are examples from an XML document after the application of Recognizer 1.

```
<s><MV><w><ph>Ochakvashe</ph><aa>Vpit+f-m2s;Vpit+f-m3s</aa>
<ta>Vpit+f-m3s</ta></w><w><ph>se</ph><aa>Ppxa---t</aa>
<ta>Ppxa---t</ta></w></MV><w><ph>rumanskite</ph><aa>A-pd</aa>
<ta>A-pd</ta></w><w><ph>firmi</ph><aa>Ncfpi</aa><ta>Ncfpi</ta></w>
<w><ph>usileno</ph><aa>Ansi;D;Vppt+cv--sni</aa><ta>D</ta></w>
<MV><w><ph>da</ph><aa>C;T</aa><ta>C</ta></w>
<w><ph>tarsjat</ph><aa>Vpit+f-r3p</aa><ta>Vpit+f-r3p</ta></w></MV>
<w><ph>dobrudzsansko</ph><aa>Ansi</aa><ta>Ansi</ta></w>
<w><ph>zarno</ph><aa>Ncnsi</aa><ta>Ncnsi</ta></w><pt>.</pt></s>

<s><w><ph>Za</ph><aa>R</aa><ta>R</ta></w>
<w><ph>momenta</ph><aa>Ncmsh;Ncmt</aa><ta>Ncmsh</ta></w>
<w><ph>kupuvachite</ph><aa>Ncmpd</aa><ta>Ncmpd</ta></w>
<XV><w><ph>sa</ph><aa>Vx---f-r3p</aa><ta>Vx---f-r3p</ta></w></XV>
<w><ph>sklonni</ph><aa>A-pi</aa><ta>A-pi</ta></w>
<MV><w><ph>da</ph><aa>C;T</aa><ta>C</ta></w>
<w><ph>brojat</ph><aa>Ncmsf;Vpit+f-r3p</aa>
<ta>Vpit+f-r3p</ta></w></MV><tok type="num">175</tok><tok
type="cyr">lv</tok><pt>.</pt></s>
```

5 Conclusion and further development

The results of the work presented in this paper are the initial stage in the development of a model for the description and processing of the group of the predicate at subsequent linguistic information levels. The model is a building block of the BulTreeBank framework and represents a component of the grammar of Bulgarian sentences. Thus the grammar that identifies and describes segments in the group of the predicate has to undergo the testing and development stages in the construction of a treebank of Bulgarian.

References

[1] Abney, S. *Syntactic Affixation and Performance Structures.* In: D. Bouchard and K. Leffel (eds.) Views of Phrase Structure. Kluwer Academic Publishers, 1990.

[2] Abney, S. *Parsing By Chunks.* In: R. Berwick, S. Abney and C. Tenny (eds.) Principle-Based Parsing, Kluwer Academic Publishers, 1991.

[3] Abney, S. *Partial parsing via finite-state cascades.* In: J. Caroll (ed.) Proceedings of the ESSLLI'96 Robust Parsing Workshop.

[4] Abney, S. *Chunk Stylebook.* Manuscript, 1996, http://www.vinartus.com/spa/publications.html

[5] Avgustinova, T. *Word Order and Clitics in Bulgarian.* Saarbruecken Dissertations, University of the Saarland, 1997.

[6] Clement, L., A. Kinyon. *Chunking, Marking and Searching a Morphosyntactically Annotated Corpus for French.* Proceedings ACIDCA'2000, Monastir.

[7] Grefenstette, G. *Light parsing as finite-state filtering.* In: Proceedings of ECAI'96, Budapest, 1996.

[8] Kinyon, A. *A Language-Independent Shallow Parser Compiler.* In: Proceedings of 10th EACL Conference, Toulouse, France, July 2001, pp.322-329.

[9] Miller, P., I. Sag. *French Clitic Movement without Clitics or Movement.* In: Natural Language and Linguistic Theory:15, 1997, pp.573–639.

[10] Monachesi, P. *Decomposing Italian Clitics.* In: Malari, S., L. Dini (eds.) Romance in HPSG, CSLI Publications, Stanford, USA, 1998, pp.305-357.

[11] Müller, F. *Shallow-parsing stylebook for German.* Technical report. Seminar fuer Sprachwissenschaft, University of Tuebingen, 2002.

[12] Simov, K., P. Osenova, M. Slavcheva, S. Kolhovska, E. Balabanova, D. Doikov, K. Ivanova, A. Simov, M. Kouylekov. *Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank.* In: Proceedings of LREC 2002, Canary Islands, Spain.

[13] Simov, K., Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, A. Kiryakov. *ClaRK - an XML-Based System for Corpora Development.* In: Proceedings of Corpus Linguistics 2001 Conference, pp.558-560

[14] Simov, K., M. Kouylekov, A. Simov. *Cascaded Regular Grammars over XML Documents.* In: Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002), Taipei, Taiwan, September 2002 (to appear).

[15] Sperberg-McQueen, C.M., L. Burnard (eds.) *TEI P4: Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen, 2002, http://www.tei-c.org/

[16] XPath, 1999. *XML Path Language (XPath) version 1.0.* W3C Recommendation. http://www.w3.org/TR/xslt