BTB-TR05: BulTreeBank Stylebook[*] BulTreeBank Version 1.0

Petya Osenova, Kiril Simov
BulTreeBank project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G.Bonchev Str. 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.org

BulTreeBank Technical Report BTB-TR05

31.08.2004

Abstract

This document presents a style book for the syntactic annotation of the Bulgarian HPSG-based Syntactic Treebank — BulTreeBank. The design approach to the syntactic annotation was to use the language model defined within a linguistic theory as a basis for the creation of an annotation scheme. In this way we ensured the consistency of the represented linguistic information, minimal human labor during the creation of the treebank and, last but not least, good validation of the treebank. Although the Treebank is theory dependant, we expect that it can be easily transformed into other formats.

1 Introduction

The Treebank of Bulgarian (BulTreeBank) is a corpus of syntactically annotated Bulgarian sentences. The syntactic annotation is based on HPSG as a language model. It consists of two sets of sentences: grammar-derived examples (1 500 sentences) and corpus-derived ones (10 000) from newspapers, government documents, prose. The source of the first set is the *Bulgarian grammars* and the source of the second set is the *electronic corpus*. Needless to say, both sub-banks of sentences represent the same linguistic phenomena, but from a slightly different perspective and with different distribution and variation. Our grammar-derived sentences are of two kinds: they have already been extracted from the Bulgarian literature to illustrate a specific phenomenon, or were constructed by the author for the same purpose. In this respect they are 'intended' sentences. Thus the annotation follows the general pre-classification as far as it is compatible with our formalized scheme. The Bulgarian grammar books that we used are as follows: Bulgarian Academy Grammar [Bulgarian Academy Grammar 1983], Contemporary Bulgarian Language — Encyclopedia (the Part on Syntax was written by Yordan Penchev) [Boyadzjiev et. al. 1999], and Bulgarian Syntax [Brezinski 2001].

Annotating corpus-derived sentences already required additional strategies for an adequate interpretation. For example, when sentences of whole paragraphs or even divisions were annotated, then a more elaborate co-reference mechanism was needed across sentence boundaries. The other thing is that the texts supply a richer variety of syntactic relations, typical for the connected text: a high frequency of introductory words (phrases), more complex word order, attachment ambiguities, dialogue patterns, nonstandard punctuation decoding. There are cases of a head-complement dependency between two sentences. In comparison with the isolated sentences, the context-bound ones are more complex for syntactic annotation, but at the same time - easier for reference and ellipsis resolution. The texts in this part of the treebank are from the following sources: Newspapers:

[*] The work reported here is done within the BulTreeBank project. The project 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.

"Novinar", Sofia; "Sega", Sofia; "Standart", Sofia; "Bulgarian Army", Sofia; Publishing houses: Nauka i izkustvo Publishing House, Sofia; Svetlostruj Publishing House, Sofia; Government documents; Bulgarian Constitution 1991; Parliament debates and documents; Laws; Web popular documents; Others (isolated paragraphs from about twenty national and regional newspapers).

In this stylebook we briefly present the basic assumptions behind the HPSG language model, which we adhered in the annotation scheme. Then the general architecture of our annotation scheme is described and exemplified. We focus on both: the syntactic domains and the linguistic phenomena. With respect to the domains, we first discuss the lexical domains, and then the phrasal ones. Functional domains (such as Pragmatic, Discontinuous etc.) are separately discussed as well. For the cases when more than one linguistically motivated decision is possible, we used some preference rules

2 HPSG Language Model

We first present the general language model, accepted within HPSG, and which we use for the creation of our treebank. HPSG is a lexicalist linguistic theory, in which the *linguistic objects* are represented via feature structures. An HPSG grammar comprises a linguistic ontology (*sort hierarchy*) and *grammar principles* (*constraints over the sort hierarchy*). The sort hierarchy represents the main types of linguistic objects and their basic characteristics. The principles impose restrictions on the objects and thus predict the well-formed phrases. A basic mechanism for ensuring the right sharing of information among the various parts of the linguistic objects is the *reentrance* (or *structure sharing*). Generally, reentrance means that a linguistic object is connected to another linguistic object via two different characteristics (represented via features or paths of features).

2.1 Linguistic Objects

Within HPSG it is accepted that linguistic objects are characterized by their unique minimal sort (see below) and a set of characteristics (features) which are partial functions that connect the objects which have these features with other objects that are values for the features. In the linguistic analyses the linguistic objects in HPSG are represented as a special kind of graphs. Here we give some very basic definitions without going into detail.

2.2 Sort Hierarchy

The sort hierarchy determines the possible kinds of linguistic objects. It defines the sorts, a relation of inheritance and the appropriate features for the sorts. The inheritance can be multi-dimensional, i.e. a sort can inherit from more than one sort. Here we will present sort hierarchies in the following format: first we represent a given sort (we use italic for sorts — *sort*, sorts are always in small letters), then at the next lines with certain leading blank symbols we declare the features appropriate for the sort (we are using bold for features — **FEAT**, features are always in capital letters). The feature declarations are of the form: **FEAT**:*sort*, where **FEAT** is the feature and *sort* is the sort of the objects that can be value for the feature of the objects of the given sort. Feature declarations can be omitted if the sort has no appropriate features of its own. Then the sub-sorts are given with leading blank symbols. The sub-hierarchy for each sub-sort is given after the sub-sort. Each sub-sort inherits the features of its super-sort in addition to its own appropriate features. When a feature is inherited it can have a more specific value sort as well. Here is an example of a sort hierarchy:

```
root
   FEAT-A : subsort1
   FEAT-B : subsort2
    subsort1
    subsort2
        FEAT-B : subsort4
        FEAT-C : subsort2
         subsort4
         subsort5
    subsort3
```

There are six sorts (*root*, *subsort1*, *subsort2*, *subsort3*, *subsort4*, *subsort5*) and three features (**FEAT-A**, **FEAT-B**, **FEAT-C**). Sort *root* has two appropriate features: **FEAT-A** and **FEAT-B** with value restrictions to sorts *subsort1* and *subsort2* respectively. The sort *root* has three sub-sorts: *subsort1*, *subsort2*, *subsort3*. Each of them inherits the two features from the sort *root*. The sort *subsort2* has one additional feature **FEAT-C** and an additional restriction over the values of the feature **FEAT-B** to sort *subsort4*. The sort *subsort2* has two sub-sorts: *subsort4*, *subsort5*. They inherit all the features from the sort *subsort2*: **FEAT-A**, **FEAT-B**, **FEAT-C**. The inheritance is a transitive relation. The sorts *subsort1*, *subsort3*, *subsort4*, and *subsort5* are minimal sorts.

The sort hierarchy plays an important role for the definition of the linguistic objects. Each object has exactly one minimal sort. If in the sort hierarchy a feature is declared as appropriate for a sort, then this feature is defined for each object of this sort and the value is of the sort given as a restriction. In this respect the sort hierarchy predetermines the possible linguistic objects. The actual linguistic objects are additionally constrained by the grammar principles.

2.3 Grammar Principles

Grammar principles are logical statements (formulas, descriptions) which are evaluated as *true* or *false* over linguistic objects. The grammar principles have the format of implications:

$$A \text{ -> } B$$

where *A* and *B* are descriptions consisting of the logical operators like negation, conjunction, disjunction, and elementary descriptions that determine the possible sorts of the linguistic object, its features and the sort of their values. The grammar principles are interpreted in the following way: each object that satisfies the description *A* has to satisfy also the description *B*. Here we will not go into any further details. The interested reader can consult the following works: [Pollard and Sag 1994], [Wasow, Bender, and Sag 2003].

2.4 Feature Graphs — Representation of Linguistic Objects

Here we give some formal definitions. They are given for completeness of the pictures, but they are not used in the report. We shortly present the syntax of the logic (SRL). For full description see [King 1989]. In [Simov 2001], [Simov et. al. 2002], [Simov 2002] we show that this normal form is suitable for the representation of an HPSG corpus and an HPSG grammar (see also [King and Simov 1998]).

Sign = <*S*,*F*,*A*> is a finite *SRL signature* iff *S* is a finite set of *species* (minimal sorts in a sort hierarchy), *F* is a set of *features*, and *A* : *S* x *F* -> *Pow*(*S*) is an *appropriateness function*.

SRL signatures represent the minimal sorts in the sort hierarchy together with all features appropriate for these sorts. Having in mind the interpretation of the sort hierarchy there is no lost of generality by using SRL signatures in the following definitions.

t is a *term* iff *t* is a member of the smallest set *T* such that (1) : is in *T*, and (2) for each *f* in *F* and each *t* in *T*, *tf* is also in *T*. *d* is a *description* iff *d* is a member of the smallest set *D* such that (1) for each *s* in *S* and for each *t* in *T*, *t* ~ *s* is in *D*, (2) for each *t1*in *T* and *t2* in *T*, *t1* = *t2* is in *D*, (3) for each *d* in *D*, *neg d* is in *D*, (4) for each *d1* in *D* and *d2* in *D*, [*d1 and d2*] is in *D*, [*d1 or d2*] is in *D*, and [*d1* -> *d2*] is in *D*. Each subset *th* of *D* is an *SRL theory*. Here we do not use the typical logical symbols, but we hope that the definitions are comprehensible.

An HPSG grammar in SRL consists of: (1) a signature *Sign*, which gives the ontology of entities that exist in the universe and the appropriateness conditions on them, and (2) a theory *th*, which gives the restrictions upon these entities.

Let *Sign* = <*S*,*F*,*A*> be a finite signature. A *feature graph* with respect to *Sign* is a directed, connected and rooted graph *G* = <*N*,*Ar*,*rn*,*SF*> such that: (1) *N* is a set of *nodes*, (2) *Ar* : *N* x *F* -> *N* is a partial *arc function*, (3) *rn* is a *root node*, (4) *SF* : *N* -> *S* is a total *species assignment function*, and (5) for each *n1*, *n2* in *N* and each *f* in *F* such that *Ar*(*n1*,*f*) is defined and *Ar*(*n1*,*f*) = *n2*, then *SF*(*n2*) is in *A*(*SF*(*n1*),*f*). We say that the feature graph *G* is *finite* if and only if the set of nodes is finite. A feature graph *G* = <*N*, *Ar*, *rn*, *SF*> such that for each node *n* in *N* and each feature *f* in *F* if *A*(*SF*(*n*),*f*) is defined then *Ar*(*n*,*f*) is defined is called a *complete feature graph*.

For finite feature graphs, we could define a translation into SRL descriptions using the correspondences between paths in the graph and terms. Thus we can interpret each finite feature graph as a description in SRL. Using the set of all finite feature graphs that subsume a given infinite feature graph, we can also define the interpretation of each infinite feature graph. So, we can call these graphs satisfiable graphs. There exists an interpretation in which they denote a non-empty set of objects. Moreover, we can define a correspondence between the finite SRL theories and the feature graphs. This representation of the theory as a set of graphs has the following very important properties:

- Each graph *G* in the set of graphs is satisfiable (for some interpretation the graph *G* denotes some objects in the interpretation), and
- Each two graphs *G1*, *G2* in the set have disjoint denotations (for each interpretation there is no object in the interpretation that is denoted by the two graphs).

These properties of the set of graphs theory representation allow for the classification of the linguistic objects with respect to the graphs. We are going to use such an algorithm for the tasks connected to the creation and usage of the corpus. Also, an inference procedure over feature graphs is developed as a composition of graphs. The procedure reflects the semantics of the corresponding SRL theory.

Thus, feature graphs are adequate for the following important scenarios: (1) **Representation of an HPSG grammar.** The construction of a graph representation of a finite theory demonstrates that using feature graphs as grammar representation does not impose any restrictions over the class of possible finite grammars in SRL. (2) **Representation of an HPSG corpus.** Each sentence in the corpus is represented as a complete feature graph. One can easily establish a correspondence between the elements of the strong generative capacity of an HPSG grammar and the complete feature graphs. Thus complete feature graphs naturally become a good representation for an HPSG corpus. (3) **Representation of the annotation scheme.** We assume that an annotation scheme over the HPSG sort hierarchy can be considered a grammar. The feature graphs of such an annotation scheme will be constrained by the lexicon, which is available to the annotators, by the principles, which are stated as a theory, and by the input sentences. As a result, all the constraints that follow logically from the above sources of information can be effectively exploited during the annotation process.

In the rest of the report we work with simpler structures than feature graphs due to several reasons: (1) there are no HPSG grammar and lexicon for Bulgarian yet; (2) the complexity of graphs make them very hard to observe and manipulate by human annotators. The actual representations, which are used in the report and in the treebank, are introduced below.

2.5 HPSG Sort Hierarchy and Principles

In this subsection we present an HPSG sort hierarchy and HPSG Principles which we used as a basis for our annotation scheme. What we present here is not a complete HPSG grammar for Bulgarian. Such an elaborate grammar is left for a future work.

2.5.1 HPSG Sort Hierarchy

Within our treebank we rely on the standard sort hierarchy of signs: the sort *sign* with sub-sorts *word* and *phrase*. It is a complex entity that is assigned two features: **PHON** (string of phonemes) and **SYNSEM** (syntactic and semantic characteristics). Further within the attribute **SYNSEM** there are three important features: **CATEGORY** (which encodes the syntactic information), **CONTENT** (which encodes the semantic information) and **CONTEXT** (which encodes the pragmatic information). The selectional force of signs is represented via three features: **ARG**(ument)**-ST**(ructure), **VAL**(ency), and **MOD**(ifier). **ARG-ST** value is a list of the arguments of a lexical item. The value of the **VAL** feature is divided into **SUBJ**(ect), **COMP**(lement) and **SPEC**(ifier) features. Each of them has a list value corresponding to the dependents of the respective kind (in our work we do not use specifiers). The values of the valency lists are connected with the **ARG-ST** list of the lexical head. The valency lists determine the realized arguments of the lexical head. The **MOD** feature determines the selection of the head by an adjunct. Its value is a *synsem* object. The constituent structure is encoded for each phrase via the attribute **DTRS**. Assigning different values to this feature, HPSG theory distinguishes between (at least) the following types of phrases — *headed-phrase* and *non-headed-phrase*. The first kind is additionally divided into *head-complement*, *head-subject*, *head-adjunct* and *head-filler*. The current hierarchy of signs is presented in the following sort hierarchy:

- *sign*
  - **PHON** : *phonlist*
  - **SYNSEM** : *synsem*
  - *word*
    - **ARG-ST** : *list-of-synsem*
  - *phrase*
    - **DTRS** : *con-struc*

Different kinds of phrases are defined on the basis of the hierarchy of sub-sorts of the sort *con-struc* (stands for *constituent structure*).

- *con-struc*
  - *headed-phrase*
    - **HEAD-DTR** : *sign*
    - **COMP-DTRS** : *list-of-phrases*
    - *head-complement*
      - **HEAD-DTR** : *word*
      - **COMP-DTRS** : *ne-list-of-phrases*
    - *head-subject*
      - **HEAD-DTR** : *phrase*
      - **SUBJ-DTR** : *phrase*
      - **COMP-DTRS** : *empty-list*
    - *head-adjunct*
      - **HEAD-DTR** : *phrase*
      - **ADJUNCT-DTR** : *phrase*
      - **COMP-DTRS** : *empty-list*
      - *head-sem-adjunct*
      - *head-pragmatic-adjunct*
    - *head-only*
    - *head-filler*
      - **HEAD-DTR** : *phrase*
      - **FILLER-DTR** : *phrase*
      - **COMP-DTRS** : *empty-list*
  - *non-headed-phrase*
    - *coordination-phrase*
      - **CONJ-DTRS** : *set(sign)*
      - **CONJUNCTION-DTR** : *word*

Each phrase of sort *headed-phrase* has a head daughter (**HEAD-DTR**) which is of sort *sign* and complement daughters (**COMP-DTRS**) which are a list of phrases. The headed-phrases are further divided into *head-complement*, *head-subject*, *head-adjunct* and *head-filler* phrases. The phrases of sort *head-complement* inherit the two features with additional constraints over the values. Thus, the head daughter for these phrases has to be a lexical sign (*word*). This restriction comes from the immediate dominance schemata defined in [Pollard and Sad 1994] which say that in a head-complement phrase the head daughter has to be lexical. Also we impose a restriction over the list of the complement daughters. It has to be non-empty (*ne-list-of-phrases*). Thus, we do not allow a head-complement phrase without complement daughters. All the other sorts of headed-phrases inherit the two features with additional restrictions: the feature **HEAD-DTR** has a value which is a phrase and the feature **COMP-DTRS** has a value empty list (*empty-list*). This restriction is necessary because these sorts of phrases cannot have complements. The features are necessary to satisfy the Valence principle (see below). Each sort has one additional feature for the non-head daughter. The restriction for the value of these features is a phrase. The phrases without a head daughter (*non-headed-phrase*) have one sub-sort: *coordination-phrase* which is for coordination phrases and it has two features **CONJ-DTRS** with value restriction to a set of objects of the sort *set(sign)* and **CONJUNCTION-DTR** which has to be a lexical sign (*word*). In our treatment of coordination we change this part of hierarchy because we realized the coordinated phrases in a flat structure.

The distinction between *head-sem-adjunct* and *head-pragmatic-adjunct* is on the basis of whether the given adjunct modifies the semantics of the head or its pragmatic nature only. An example of a pragmatic adjunct are the vocative phrases in Bulgarian (see [Osenova and Simov 2002] for details).

Another part of the HPSG sort hierarchy which plays an important role in the design of our annotation scheme is the sub-hierarchy of the sort *head*. This part of the sort hierarchy is responsible for determining of the grammar features of the sign. We can consider them as generalized parts of speech.

```
head
     substantive (subst)
          PRD : boolean
          MOD : mod-synsem     none for not adjuncts
            noun
                 CASE : case
            verb
                 VFORM : vform
                 AUX : boolean
                 INV : boolean
            adj
            adv
            prep
                 PFORM : pform
```

Here we focus on the sub-sorts of the sort *subst*. They determine the main domains of phrases like noun phrases, verbal phrases, adjectival phrases, adverbial phrases and prepositional phrases. All of them inherit the two features of *subst*: **PRD** which determines whether the phrase is used predicatively or not and **MOD** which determines the selection potential of the adjuncts. In our treebank we do not have any functional heads, and for that reason the corresponding part of the sort hierarchy is missing.

The other parts of the HPSG sort hierarchy do not play such an important role in the current development of our treebank and we will not present them here. We will introduce some additional features and sorts when it is necessary, but we will not give a complete hierarchy. Sometimes it is not possible to present such a hierarchy because our analyses of Bulgarian are still not detailed and explicit enough.

2.5.2 HPSG Principles

The principles in HPSG are the mechanism for representation of the grammar. Here we present some of the basic principles which guide the linguistic analyses in the treebank.

Head Feature Principle (HFP):
The **HEAD** value of any headed phrase is structure-shared with the **HEAD** value of the head daughter.

This principle ensures for each headed phrase the right propagation of the **HEAD** value from the head daughter to the headed phrase itself.

Valence Principle (VALP):
Any valency feature **VF** value (in our case subject feature (**SUBJ**) and complement feature (**COMP**) of any headed phrase is the **VF** value of the head daughter minus all **VF** daughters.

The Valence principle governs the actual realization of the arguments of the lexical head. First, all complements (if any) are realized. This realization follows from the Valence principle and the sort hierarchy which says that in a *head-complement* phrase the head daughter is lexical. Then the subject (if any) is realized. The realization of the adjuncts is governed by the next principle.

Head-Adjunct Principle (AdjunctSP):
For each phrase with **DTRS** value of sort *head-adjunct* the **MOD** value of the adjunct daughter is the same with the **SYNSEM** value of the head daughter.

This above principle states that the adjunct daughter selects its head via the **MOD** feature.

> **Semantic Principle (SemP):**
> The **CONTENT** value of any headed phrase is structure-shared with that of the adjunct daughter if the **DTRS** value is of sort *head-sem-adjunct*, and with that of the head daughter otherwise.

This principle determines the semantics of the headed phrase. Note that it is determined by the adjunct if the adjunct is a semantic one and by the head daughter otherwise. We encode very little semantic information in the current version of the treebank, but for us the principle is important with respect to the pragmatic adjuncts which contribute to the **CONTEXT** value of the phrase and only indirectly have semantic contribution. For an example of pragmatic adjuncts see [Osenova and Simov 2002].

The last principle we present here is responsible for the realization of fillers in *head-filler* phrases:

> **Head-Filler Principle (FillerSP):**
> For each phrase with **DTRS** value of sort *head-filler* the **SLASH** value of the head daughter contains the **LOCAL** value of the filler daughter and **SLASH** value of the phrase is the same as that value of the head daughter minus the **LOCAL** value of the filler daughter.

This is a very informal definition of the Head-Filler Principle, but it is enough for the definition of our annotation schema. The general idea is that for the constituents that are not realized locally, the information is collected in the value of the **SLASH** feature and it is used to select the appropriate filler later in the structure.

We used the above mentioned elements of the HPSG theory in order to design our annotation scheme for the Bulgarian treebank. As we have limited resources for the implementation of a complete HPSG grammar for Bulgarian, we consider these elements to be a good basis for the main design principles behind the annotation scheme.

3 Annotation Scheme Based on HPSG Language Model

In order to reflect the above stated (partially) HPSG theory we designed an annotation scheme which encodes the following information: **constituency** — the graphs defined on the basis of mother-daughters relation in the constituent structure of signs is the basic representation of the sentences analysis. We select these graphs because they are close to the traditional context free tree representation. Although the graphs are not always trees we frequently (and inexact) will refer to them as trees; **category** — each sign is represented as a node in the tree with a phrasal or lexical label corresponding to the category of the sign; **head-dependent relation** — the labels of the nodes in the tree reflect the sort of the constituent structure of the phrase (*head-complement*, *head-subject*, *head-adjunct*, and *head-filler*); **linear order** — the original word order is preserved. The clashes between the word order and the constituent structure are solved by introducing discontinuous elements where necessary; **co-referential relations** — each non-inferable co-referential relation is stated explicitly (but some infereable ones are represented explicitly as well); **unexpressed elements** — those include unexpressed subject and ellipsis. The unexpressed subject and unexpressed complements are represented explicitly when necessary, i.e. they participate in a co-referential relation. Ellipsis is always represented explicitly. Pro-dropness is a characteristic feature of Bulgarian and when involved in a co-reference relation, it is represented as a *pro-ss* element.

In each headed phrase the head daughter is represented implicitly and in most cases can be inferred automatically. The mechanism of co-reference is used for phenomena like pro-dropness, secondary predication, binding etc. Some features of the signs are represented as nodes in the trees. Such nodes dominate the sign nodes for which they express some feature(s). Some other features of the signs are represented as attribute-value pairs in the XML representation of the trees. Thus in the annotation scheme the following types of elements have been distinguished:

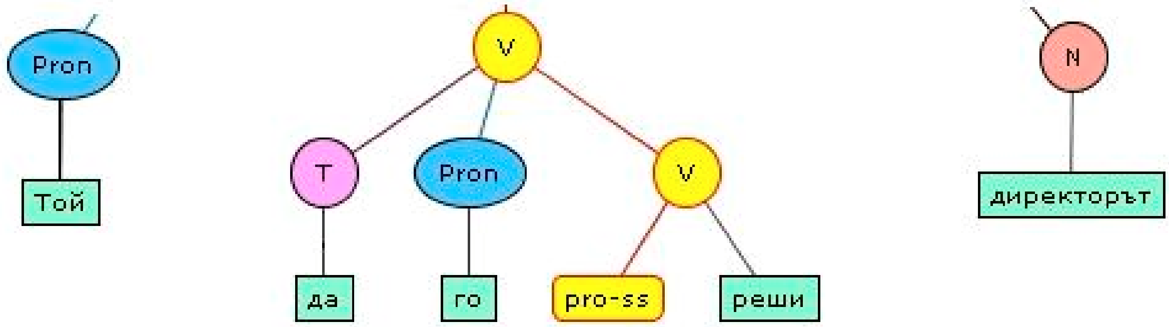
- *Lexical elements* which cover lexical signs. They include not only the lexical items in the lexicon but also the analytical word forms resulting from the application of some lexical rules. Such elements are *N*, *V*, *Prep*. The actual strings are given as elements under the lexical elements;
- *Phrasal elements* which represent: a hierarchy of phrase types like head-complement, head-adjunct and the corresponding syntactic domains like nominal, verbal phrases (we have elements like *VPA*(djunct), *NPC*(omplement));

- *Functional elements* which help us to represent the multi-dimensional nature of the encoded linguistic knowledge (thus, in order to preserve the original word order we have introduced discontinuous elements like *Disc*(ontinous)*E*(xtracted), pragmatic one *Pragmatic*, sentence's root *S*, conjunctions and arguments of coordination phrase: *Conj*, *ConjArg*, etc.;
- *Textual elements* represent the actual strings of the lexical elements and the punctuation.

These elements are described in detail in the next sections. Here we show how the different kinds of elements are depicted in the graphical representation of sentence analyses throughout this report. The graphical representations are generated from the XML representation of the trees within the CLaRK System.

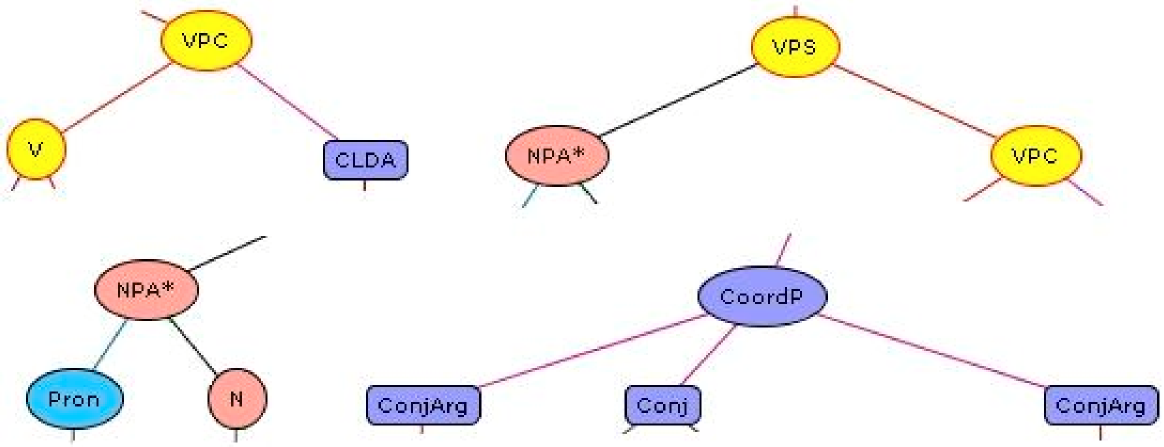
The lexical and phrasal elements are represented in the graphical view as ellipses with their labels inside. The lexical elements differ from the phrasal elements by their labels. We define different colors for the different categories of elements, thus the reader can navigate easily the layout. The head daughter of each phrase is defined implicitly. It can be recognized on the basis of the type of the phrase. For more details consult the sections below. Here are some lexical and phrasal elements:

Lexical elements



In the pictures above there are several lexical elements: two *Pron* elements, one *T* element, two *V* elements, and one *N* element. As it can be seen some lexical elements have a nested, complex structure. In this case one of the verb elements represents an analytical verb form consisting of a particle (*T*), pronoun (*Pron*) and a verb (*V*). The nested verb has also a *pro-ss* element as a child. The lexical elements have their wordforms as children.

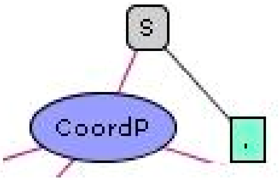
Phrasal elements

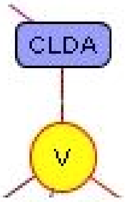


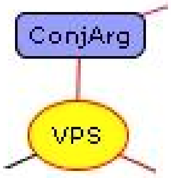
The children of a phrasal element correspond to the constituent structure of the phrasal element. The children can be other phrasal elements, lexical elements, functional elements and punctuation marks. Here we have a *VPC* element for a verb phrase of sort *head-complement* with one *CLDA* complement; a *VPS* element for a verb phrase of sort *head-subject* with a *VPC* head and an *NPA* subject; a *NPA* element for a noun phrase of sort *head-*

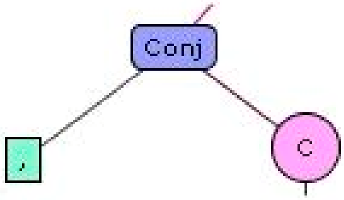
adjunct. The asterisk after the label means that this element is an argument of a co-referential relation; a *CoordP* element for a coordination phrase. It has only functional children which mark-up the role of each constituent: an argument of the coordination or a conjunction.

Functional elements are represented in the graphical view as rounded rectangles with their labels inside. Functional elements have only one child, the element to which they assign some property. Sometimes punctuation marks are also attached to functional elements. This holds especially for the root of the sentence. The end punctuation mark is always attached to the element *S*. Here are some examples:





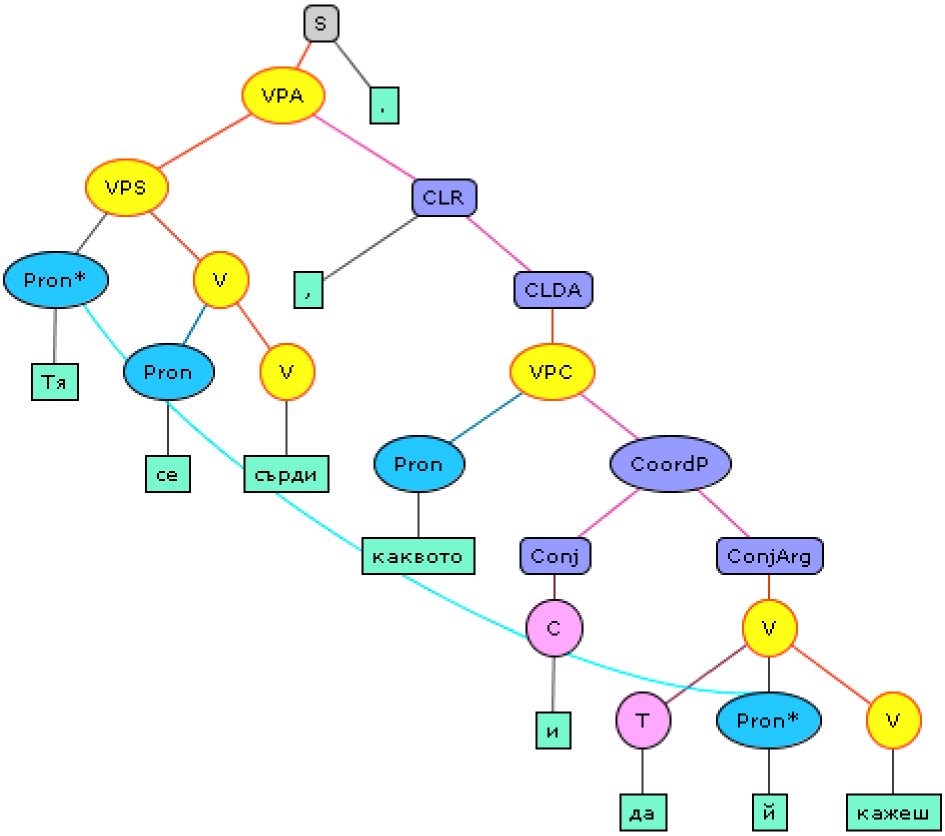




Here we have an *S* element with a child *CoordP* element and a full stop; a *CLDA* element which marks-up a clause - in this case it is an analitical lexical verb; a *ConjArg* element which represents an argument of a coordination phrase; a *Conj* element which marks-up the conjunction of a coordinated phrase together with the obligatory comma for the conjunction, if required.

Textual elements are represented in the graphical view as rectangles with their textual value inside. As it was mentioned above, they represent the actual string of the sentences including the word forms and punctuation marks. They are always leaves in the tree. For examples of textual elements see the fragments of trees given above or the following tree.

Here is an example of a complete tree in a graphics view. Keep in mind that not all information represented in the XML format of the tree is given in the graphical view. The graphical view is meant for a better illustration.



The tree represents the analysis for the sentence: 'Тя се сърди, каквото и да й кажеш.' (She is angry whatever you tell her). The textual elements represent eight word forms and two punctuation marks. Above each word form there is a lexical element with an appropriate label. The label for each lexical element is detemined on the bisis of the morpho-syntactic tag for the word form. There are also two analytical word forms 'се сърди' (a verb with a non-referential reflexive pronoun) and 'да й кажеш' (special Bulgarian 'da' verb form). For a discussion on the analytical word forms see below. Then we have four phrasal elements: *CoordP*, *VPC*, *VPS* and *VPA*. The coordination phrase (*CoordP*) comprises a conjunction (marked-up additionally with the functional element *Conj*) and an argument (marked-up additionally with the functional element *ConjArg*). This coordination is a head daughter of the *VPC* verbal phrase. The complement is a relative pronoun. The *VPC* phrase is saturated and it is a clause in the sentence. This clause has two properties: it is a special 'da'-clause because the head verb is 'da' verb form and at the same time it is a relative clause because of the linking relative pronoun. These two properties are annotated respectively with the functional elements — *CLDA* and *CLR*. The *VPS* phrase has two daughters - a head daughter *V* and a subject daughter *Pron*. The whole sentence is a *VPA* phrase with a head daughter *VPS* and an adjunct daughter *CLR*. The functional element *S* represents the root of the sentence. The two punctuation marks are attached to the two elements that they define in the sentence: comma for the relative clause and full stop for the sentence. Also, in the picture we have a representation of the coreferencial relation between the subject of the matrix sentence and the dative clitic in the verbal form of the clause. For each coreferencial relation we add an asterisk to the label of the elements that are arguments of the relation and then we connect the elements with line(s).

Here is the same tree as it is encoded within XML format in the treebank. The root element for each sentence is the element <*s*> which has two children — <*text*> and <*analysis*> (Some sentences additionally have two more children: one for the classification in the grammar book where the sentence was taken from and one to describe the source of the sentence. We will not discuss these elements here.) The element <*text*> contains the textual representation of the sentence. The element <*analysis*> contains the actual syntactic annotation. It has one or more children. Each child is an element <*S*> which contains one syntactic structure for the sentence. If the sentence has more than one reading, then there is more than one <*S*> element for it. The <*S*> element has at least three children. The first (<*Discourse*>) is reserved for the representation of the discourse relations and at the moment it is not systematically used. Thus we will not discuss it here. The second child (<*CoIndex*>) is used to represent different (non-hierarchical) relations within the sentences: discontinuity, elliptical elements, co-referential relations.

```
<s>
 <text>Тя се сърди, каквото и да й кажеш.</text>
 <analysis>
 <S>
  <Discourse>
    <InDiscourse></InDiscourse><OutDiscourse></OutDiscourse>
  </Discourse>
  <CoIndex>
    <identifier id="id6330"></identifier>
  </CoIndex>
  <VPA>
    <VPS>
      <Pron idref="id6330"><w aa="Ppe-os3f" ana="Ppe-os3f">Тя</w></Pron>
      <V>
        <Pron ref="no"><w aa="Ppxta" ana="Ppxta">се</w></Pron>
        <V><w aa="Vpitf-o2s;Vpitf-o3s;Vpitf-r3s;Vpitz--2s" ana="Vpitf-r3s">сърди</w></V>
      </V>
    </VPS>
    <CLR>
      <pt>,</pt>
      <CLDA>
        <VPC>
          <Pron><w aa="Pra--s-n;Pre--s" ana="Pre--s">каквото</w></Pron>
          <CoordP>
```

```
            <Conj>
              <C><w aa="Cp" ana="Cp">и</w></C>
            </Conj>
            <ConjArg>
               <V>
                 <T><w aa="Ta;Tx" ana="Tx">да</w></T>
                 <Pron idref="id6330">
                    <w aa="Ppetds3f;Ppetss3f;Psot--3--f" ana="Ppetds3f">й</w>
                 </Pron>
                 <V><w aa="Vpptf-r2s" ana="Vpptf-r2s">кажеш</w></V>
               </V>
            </ConjArg>
          </CoordP>
        </VPC>
      </CLDA>
    </CLR>
  </VPA>
  <pt>.</pt>
 </S>
 </analysis>
</s>
```

The next elements within <*S*> represent the syntactic analysis. If the sentence has an ending punctuation mark, it is represented on this level (in the example it is element <pt>.</pt>). The constituent structure of the sentence is represented (as much as possible) by the tree structure of the XML document. The co-referent relation in the example is represented via sharing of the same id value (id6330) by the nodes that are in the relation. For relations like *member-of* and *subset-of* additional elements are introduced.

The actual realization of the head dependents is governed by a set of immediate dominance schemata. The realization of the dependents follows the sequence: *complements* -> *subject* -> *adjunct*. This principle entails the following: if the phrase is assigned at the highest level *VPA*, it means that either the complements and the subject are already realized, or they are not explicitly expressed. It is possible, however, that other adjuncts are in order to be taken. If the phrase is *VPS*, it means that there were no adjuncts and the complements either were alredy realized, or they were not expressed. *VPC* at the highest level means that the subject is not expressed and there are no adjuncts. The actual number and kind of dependents is determined by lexical elements within each phrase. Note that each of the three phrases (*VPC*, *VPS* and *VPA*) can be missing in the sentence structure depending on the possibility to express or not to express a certain grammatical role. In the order: *complements* -> *subject* -> *adjuncts*, we say that a subject is a higher dependent than a complement, and an adjunct is a higher dependent than a subject and a complement. In the opposite direction a complement is a lower dependent than a subject and/or an adjunct, a subject is a lower dependent than an adjunct. All this is within the projection of the same lexical head.

Thus our general assumptions are:

- Constituent structure is separated from the linear order.
- The elements of the constituent structure are not coming in any particular order.
- The elements of a constituent are usually realized adjacently (continuously). We do not classify the different orders in some special way. Thus, all acceptable permutations are treated as the same structure:

  (1) Мъжът   целува   момичето
  man-the[nom]   is-kissing girl-the[acc]
  the man is kissing the girl.

  (2) Целува   момичето   мъжът
  is-kissing girl-the[acc]   man-the[nom]
  the man is kissing the girl.

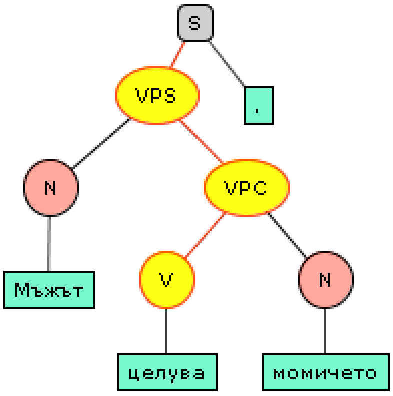
  (3) Момичето   целува   мъжът
  girl-the[acc]   is-kissing man-the[nom]
  the man is kissing the girl.

(4) Мъжът momичето целува.
man-the[nom] girl-the[acc] is-kissing
the man is kissing the girl.

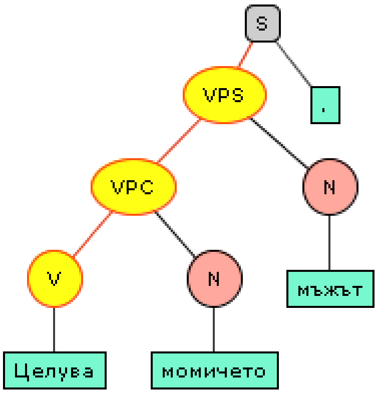
- Each constituent represents a grammatical function. This is an argument against the flat structure.
- The category of the head of the phrase determines the (traditional) domain (type) of the phrase.

Here we present the graphical view of the four combinations of the different word orders above:

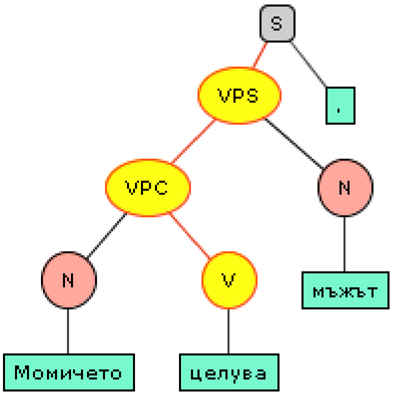
(1)



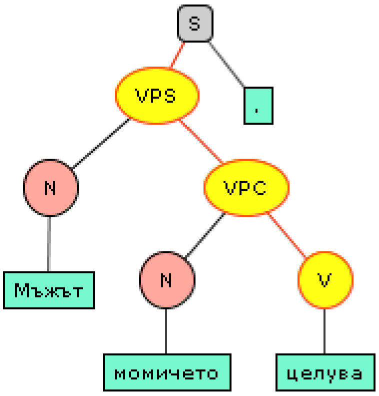
(2)



(3)

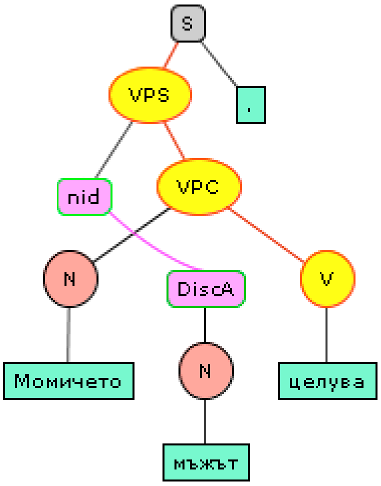


(4)



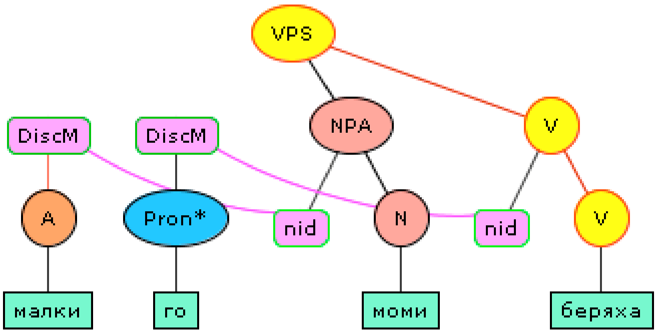
If the elements of a constituent are not realized adjacently, then discontinuous constituents are introduced. There are three types of discontinuous constituents, which we distinguish on the basis of the head properties:

Head dependents permutation (*DiscA* functional element). A higher dependent is realized between the head and a lower dependent(s). It means that there is one head and this head has at least two swaped, but adjacent dependants. Here is an example in which the subject is realized between the lexical head and the complement.

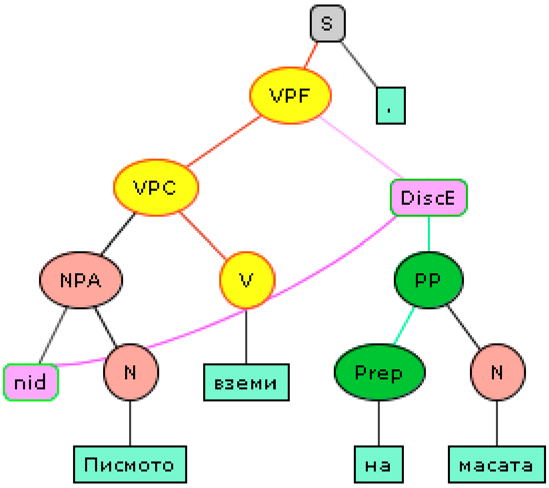


In order to represent the linear order we mark-up the higher constituent with the functional element *DiscA* and annotate it at the higher place with the functional element *nid* (stands for *non immediate dominance*) and connect the two functional elements with a line. The connection between the *DiscA* element and its immediate dominance element (*VPC* in the example) is not shown in the graphical view. This demonstrates that the element in fact belongs to another constituent.

Mixture of two constituents (*DiscM* functional element). The elements of two constituent structures are mixed without any one of them to be the governor of the other. It means that there are at least two heads and respectively, two dependents. Below (on the next page) there is an example in which the verbal clitic 'го' (it) is placed between the modifing adjective and its head noun. Again, in order to represent the linear order we mark-up the misplaced constituents with the functional element *DiscM* and annotate their real positions with the functional element *nid* (stands for *non immediate dominace*), and connect the two functional elements with lines. The connections between the *DiscM* elements and their immediate dominance elements (*VPS* in the example) are not shown in the graphical view. This is done in order to demonstrate that the elements in fact belong to other constituents.



External realization of an inner constituent (*DiscE* functional element). This is the case referred to generally as *extraction*. It means that the element is governed by a lower positioned head. Here is one example of an extracted element. The dependent PP is extracted from the noun phrase. The extracted element is annotated with *DiscE* element. Such an element is always a daughter of a *VPF* element (there are exceptions — see below), because its realization is governed by a *head-filler* phrase. This is why there is a line between the *DiscE* element and the *VPF* element.



One important point here is that the *nid* is a mechanism for showing that a misplaced constituent belongs to a certain phrase but it is not a mechanism for showing the word order of the misplaced element within the constituent to which it belongs. Thus, we always present the *nid* element as a first daughter.

In the next sections we present the lexical and phrasal domains in parallel to the language phenomena that are covered by our treebank.

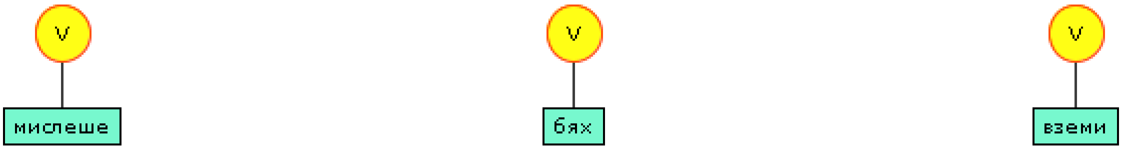
3 Lexical Elements

Lexical elements correspond to lexical signs (sort *word*). The annotation of various tokens is described in [Simov and Osenova 2004b]. The lexical elements introduced here aim at a uniform treatment of the word forms with different morpho-syntactic characteristics. Lexical elements include *isolated word forms, multi-token word forms and analytical word forms*. Here we first present the mapping from the morpho-syntactic tags to the lexical elements and then we discuss some of their special properties.

Lexical elements have a lexical category: *V* (verb), *Participle* (participle), *Gerund* (gerund), *N* (noun), *Pron* (pronoun), *A* (adjective), *M* (numeral), *H* (family name or adjective, derived from family names), *Adv* (adverb), *C* (conjunction), *Prep* (preposition), *T* (particle), *I* (interjection).

3.1 Verbal Lexical Element

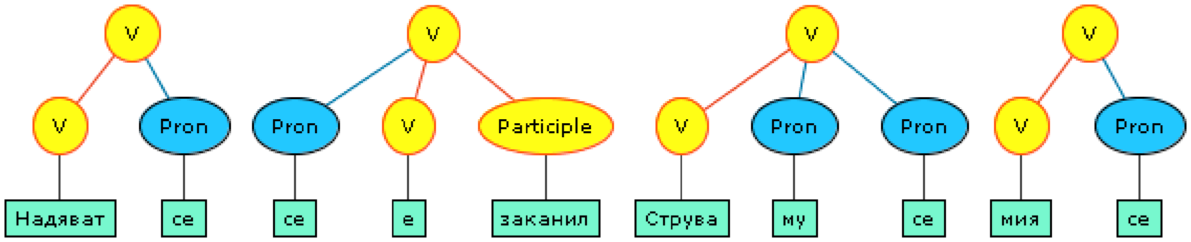
Verbal lexical element (*V*) corresponds to a single verb, verb particle or to an analytical verb phrase (passive forms, mood, tenses etc.). The single word form is annotated with the lexical element *V* if its morpho-syntactic tag maps the following pattern: "**V%%%f#**", "**V%%%z#**", or "**Tv**" where **V** maps to the letter V, **T** maps to the letter T, **f** maps to the letter f, **v** maps to v, **z** maps to z, **%** maps to any letter, **#** maps to any sequence of letters. This means that a single word form is annotated with *V* when it is a personal verb in indicative or imperative mood. In the graphical view we represent such word forms in the following way:



As multi-token verbs we consider the verbal complex in which the head is a finite verb accompanied by clitics, auxiliary particles, emphatic adverbs and participles. Here we present a list of the different cases.

Verbal Complex with Reflexive Accusative and Dative clitics

The short accusative (morphological tag *Ppxla*) and dative (morphological tag *Ppxld*) clitics mark both: the real reflexives and the non-reflexive usages, in which the clitics are considered part of the verbal lexeme like in the verb 'усмихвам се' (to smile). On the other hand, these clitics do not have a fixed position with respect to the verb itself. They can appear after the verb, in front of it, or even be separated from it by another wordform. The last case is limited to the auxiliary verb form 'е' ('съм' (to be) in third person, singular) in perfect verb tense with a participle or some of the other verbal clitics. Here are several examples:

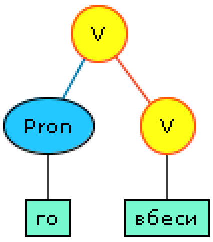


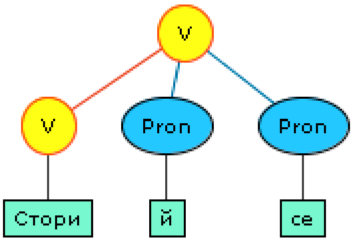
The first example shows the clitic as an enclitic in second position; the second example shows a remote realization of the clitic when the auxiliary verb is between the clitic and the participle; the third example demonstrates a realization of the dative non-reflexive clitic between the verb and the reflexive clitic. The fourth example illustrates a verb with a reflexive clitic, which also has a reflexive meaning (in contrast to the other three examples, where the clitic is reflexive only by its form).

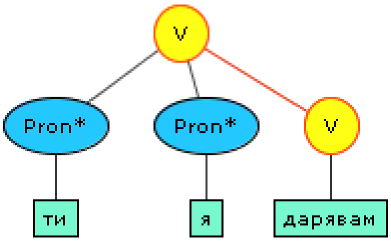
In order to represent the fact that sometimes the reflexive clitics (also some of the other pronouns) are not referring to an entity, we introduce the attribute *ref* for the element *Pron* which by default has the value *yes* for the cases when the pronoun is a referring one and value *no* for the cases when it is not a referring one. This attribute is presented in the XML encoding of the treebank and it is not shown in the graphical view.

Verbal Complex with Non-reflexive Accusative and Dative clitics

These clitics correspond to the direct (accusative) and indirect (dative) objects of the verb. In our treatment of the clitics they are realized lexically without occupying a complement position. These clitics change in number, person and gender (third person, singular). The clitics are: accusative: 'ме' (*me short form*), 'те' (*you short form*), 'го' (*him short form*), 'я' (*her short form*), 'го' (*it short form*), 'ни' (*us short form*), 'ви' (*you short form*), 'ги' (*them short form*); dative: 'ми' (*me short form*), 'ти' (*you short form*), 'му' (*him short form*), 'й' (*her short form*), 'му' (*it short form*), 'ни' (*us short form*), 'ви' (*you short form*), 'им' (*them short form*). Both clitics can be realized simultaneously. In this case the dative clitic is always positioned before the accusative one. The grammatical information from the clitics is added to the **ARG-ST** list of the verb and if there are full fledged complements of the verb, they have to agree with the clitics in number, gender and person. Below there are some examples. In the first example we have an accusative clitic realized before the verb. In the second example a dative clitic is realized after the verb, but before the reflexive clitic. The third example demonstrates the realization of both kinds of clitics — first the dative clitic, then the accusative one and last comes the verb.

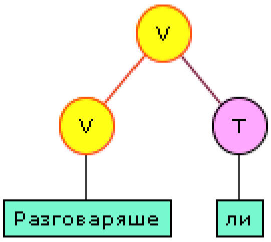


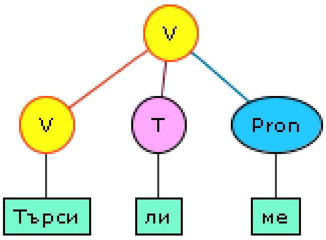


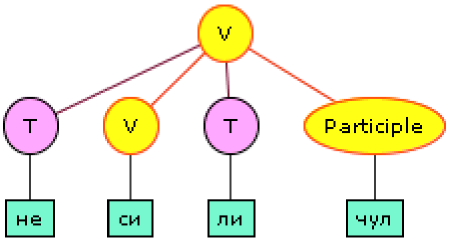


Verbal Complex with an Interrogative clitic —the interrogative particle 'ли'

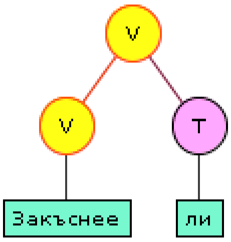
The interrogative particle 'ли' in the verbal complex is always realized adjacent to the verb, immediately after or before it. It can play two roles. First, it is a marker of interrogativity. In this case it converts the sentence (clause) into an interrogative one. Second, it can be a marker of condition (synonymic to ако 'if'). In this case it is attached to a perfective verb. If the verbal form contains an auxiliary and a participle the clitic is attached to the participle. Here are some examples:

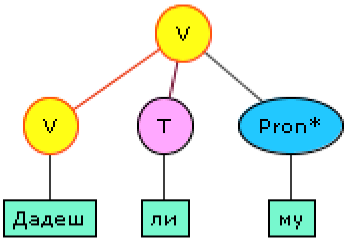






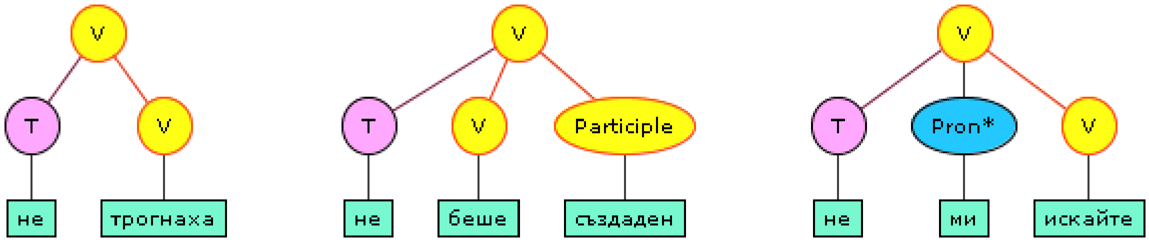
In the above examples the interrogative particle (*T*) is a question marker. The second and the third examples demonstrate the adjacent realization of this clitic to the verb when used with other clitics (participle in the third example). Below there are some examples for the second role of the particle:





Verbal Complex with a Negative clitic — the negative particle 'не'

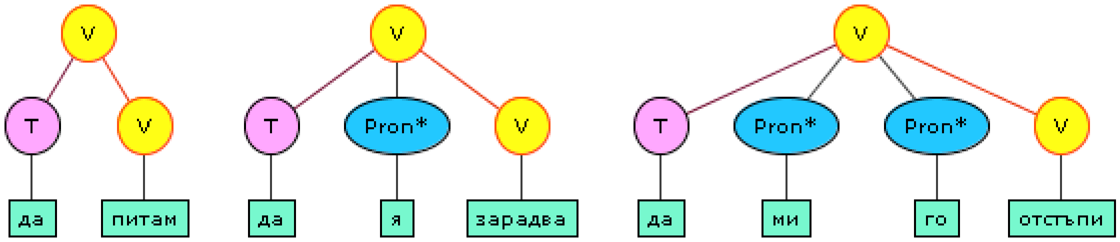
The negative particle 'не' is always realized first in the verbal complex except for the case of presence of the auxiliary particle 'да' (see below). It negates the whole verbal complex and then the sentence. Here are some examples:



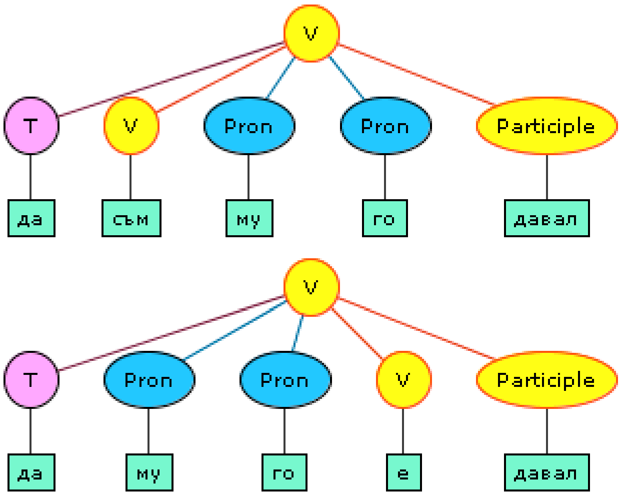
The first example shows the simplest case of a negative particle and a verb. The second example presents the negative particle in a verb form consisting of an auxiliary and a participle. The third example demonstrates the interaction of ‘не’ with the other clitics in the verbal complex.

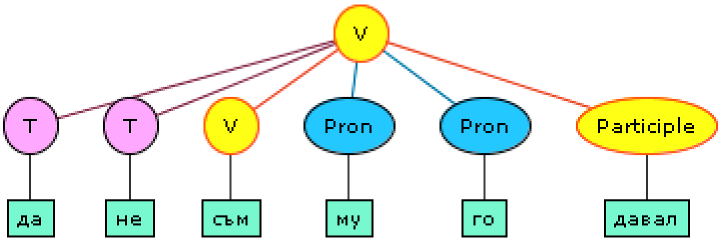
Verbal Complex with the Auxiliary particle 'да'

The auxiliary particle 'да' (to) selects a verb in present tense and forms a verbal complex which inherits the **ARG-ST** list of the selected verb and which is tenseless. Such a verb form is the lexical head of the so called 'da'-clause in Bulgarian which will be discussed later in the report. It is similar to the infinitival clause in English. All the clitics are realized between the auxiliary particle and the verb. In perfect tense the accusative and dative clitics can be realized between the auxiliary particle and the auxiliary verb, or between the auxiliary verb and the participle. Here are some examples:



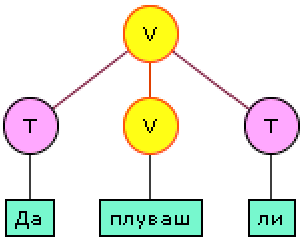
The first example above shows the simplest case of auxiliary particle 'да' and a verb. The second example demonstrates additionally an accusative clitic between the auxiliary particle and the verb. The third example presents the case of an accusative and a dative clitics within such a kind of verbal complex. The next examples display the realization of the clitics in case of auxiliary verb and a participle (perfect tense):

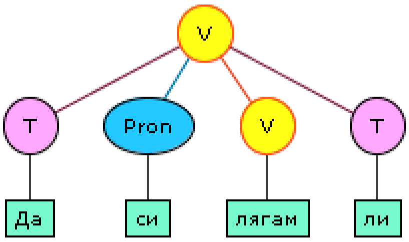


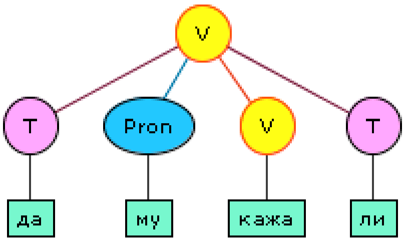


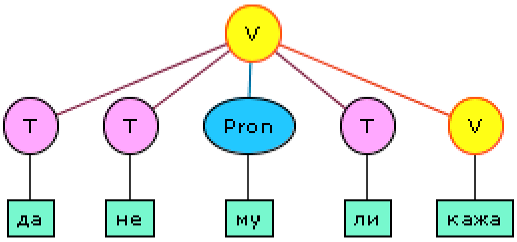
The first example above shows the realization of the clitics between the auxiliary verb and participle. The second example demonstrates another realization. The difference is only when the auxiliary verb is third person, singular. Then it comes next to the participle. The negative particle is always realized immediately after the auxiliary particle ‘da’ as it is shown in the third example.

The interrogative particle is realized just after or before the verb or the participle as in the following examples. The second example shows the realization together with a reflexive clitic. The third example shows also a dative clitic. The last example demonstrates the realization of the interrogative clitic before the verb.





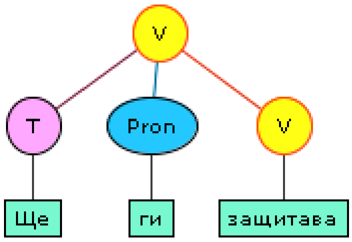


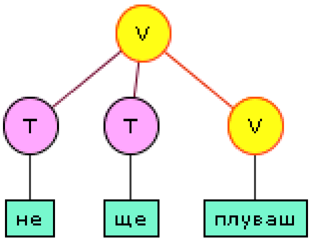


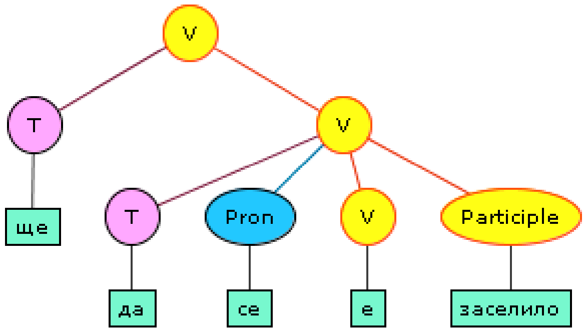
The auxiliary particle 'да' can also select verbs in pluskvamperfect and imperfect tense. In both cases the verbal complex can express different modal varieties like conditional event and others. These verbal forms follow the patterns of realization of the clitics above, if any. This is why we will not give such examples.

Verbal Complex with the Auxiliary particle 'ще'

The auxiliary particle 'ще' also plays several roles in the formation of the verbal complex. The first and primary role is the formation of future tense and future perfect. Here we first discuss this role and then we present also the other roles of the particle. In many respects this particle is similar to the auxiliary particle 'да', but there are some differences. The auxiliary particle 'ще' selects for a main verb in present tense and forms future tense of the verb. It can select verb forms comprising an auxiliary verb 'съм' (to be) and an aorist participle (Future Perfect). The realization of the clitics is the same as in the case of the auxiliary particle 'да' except for the negative particle. The difference is that the negative particle 'не' is realized before the auxiliary particle 'ще'. This negative form of the future tense is rare. In most cases the negative future form or the negative future perfect is expressed with the help of 'da'-clause selected by the forms of the negative verb няма. See the section on the verbal phrases below. Here we give some examples of future tense verb forms. The first example shows a future form with an accusative clitic. The second example demonstrates the negation of the future form. The last example demostrates the second role of the auxiliary particle 'ще'. It is to form the so-called presumptive form: 'ще да е отишъл' (he might have gone). In this case the auxiliary particle 'ще' selects for a 'da' verbal form. We annotate first the 'da' form and then the presumptive. Note that this differentiation is done within the lexical sign.







.

Analytical verb forms with the auxiliary verbs 'съм', 'бъда', 'бивам'

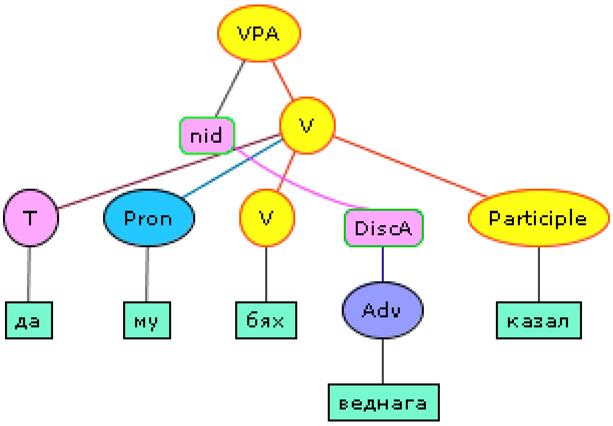
A great part of the Bulgarian analytical verb forms are constructed as a combination of forms of the auxiliary verbs 'съм', 'бъда', 'бивам' and aorist, imperfect or passive participle. Here we are not going to present the full picture of this large set of forms. Such forms already were presented in the above examples.

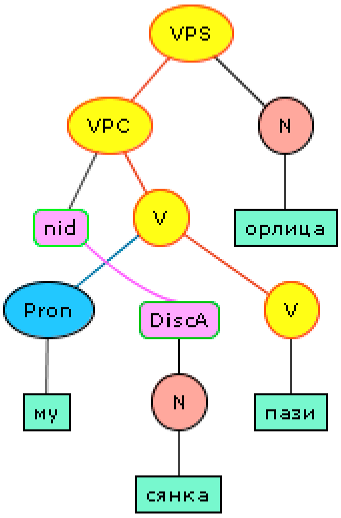
Analytical verb forms with the verb 'ща'

According to the traditional Bulgarian grammar a part of the Bulgarian verb paradigm is constructed with the past forms of the verb 'ща' and 'da'-form of the main verb. In BulTreeBank we always consider the 'da' form as a lexical head of a clause. Thus the analytical forms of this kind are considered as a main verb 'ща' which selects for a 'da' clause. For more details on this see the section of verb phrases below.

Discontinuity of analytical verb forms

The element of an analytical verb form can be separated by external material. Usually the point of discontinuity is between the auxiliary verb and the participle, but in some special kinds of text other elements can be separated (these cases are rare). Here we present two examples:



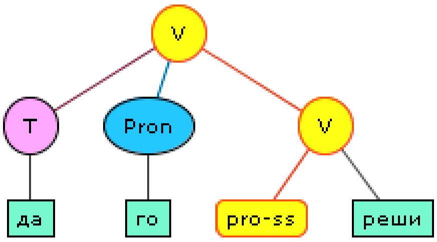


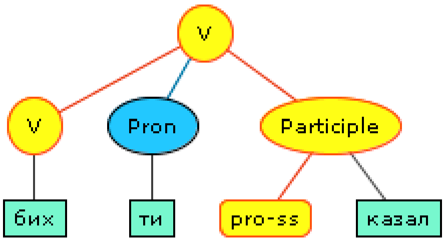
The first example demonstrates the most frequent case when an adverb or adverbial phrase is between the auxiliary verb and the participle. The second example shows a rare case when a complement is realized inside the verbal complex.

The verbal complex can also be separated by pragmatic constituents like parenthetical expressions, but they are not marked as discontinuous.

Unexpressed subject

As it was already mentioned, pro-dropness is a characteristic feature of Bulgarian. We annotate the unexpressed subject only when it participates in a co-referential relation. Then it is represented as a *pro-ss* element. This element is attached to the main verb in the verb complex (sometimes it is represented under a participle). Note that the *pro-ss* element does not show the word order position of the subject. When participles are used as nominal modifiers, they do not have a *pro-ss* element and it is accepted that their subject is coreferred with the head noun. Here are some examples:





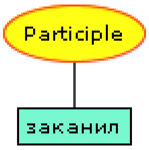
Other peculiarities

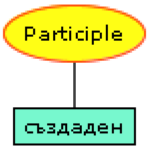
In addition to personal and participle forms as a main verb in a verbal complex we can have also a coordination, verbalised form and elliptical verb forms. The coordination is annotated as a *CoordP* phrase (see below) where the arguments of the coordination are lexical verbs. The elliptical verb forms are annotated as *V-Elip* or *VD-Elip* elements. In cases of mixed discountinuity a *V* element can have a *nid* element representing the non-local element of the verb complex. This is a rare phenomenon. For description of the mixed dicountinuity see above. A verb can be modified by the so-called emphasising words like 'само' (only) and the result is a verb again. We do not annotate the participle as a *V* element even if it is a lexical head for a sentence or a clause.

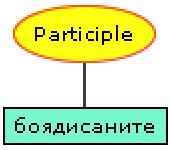
As verb lexical element we also consider periphrastic constructions of the type: 'имам направени' ('have-I made', I have something made) which equals 'направил съм' (have made-I, I have made).

3.2 Participle Lexical Element

Participle lexical element (*Praticiple*) corresponds to a single participle or to an analytical participle phrase. The single word form is annotated with the lexical element *Participle* if its morpho-syntactic tag maps the following pattern: "**V%%%c#**" where **V** maps to the letter V, **c** maps to the letter c, **%** maps to any letter, **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:

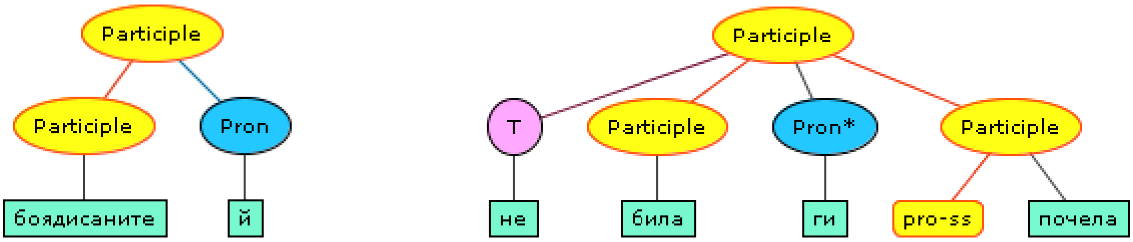






Having the same selectional potential as the verb lexeme and sharing some features with nominals, each participle can have clitics (reflexive, not-reflexive — accusative and dative, interrogative, negative, and possessive clitics in attributively used participles). The negative particle is realized as a part of the word form when the participle is used attributively in a noun phrase. There are analytical participle forms which are constituted by a participle of the auxiliary verb and a participle of the main verb. As it was explained above the unexpressed subject can be represented as a *pro-ss* element in cases when the participle is used as a lexical head in a sentence or a clause. In cases when the participle modifies a noun it has no *pro-ss* element. The *pro-ss* element does not show the word order of the subject. Participles can be conjoined to form lexical coordination

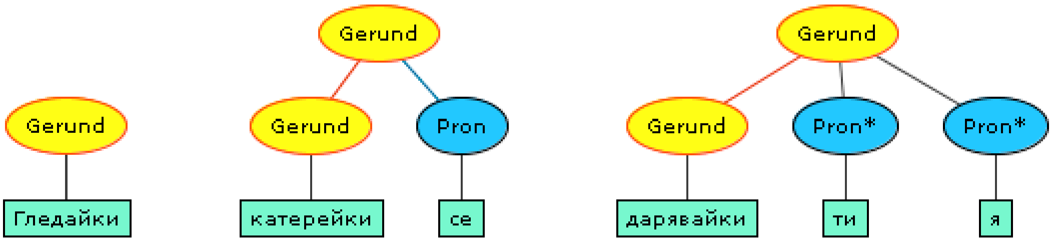
(*CoordP* element) which is also considered as a participle. Complex participle phrases can be discontinuous. Usually an adverb or adverbial phrase may appear between the auxiliary participle and the main participle. A participle can be modified by the so-called emphasising words like 'само' (only) and the result is a participle again. Here we present some examples of complex participles:



The first example shows a participle, which can be used only attributively because it is definite and which has a possessive clitic. The second example is an analytical participle form which includes also a negative particle and an accusative clitic. As it can be seen from the example, the order of the clitics is the same as for the *V* element.

3.3 Gerund Lexical Element

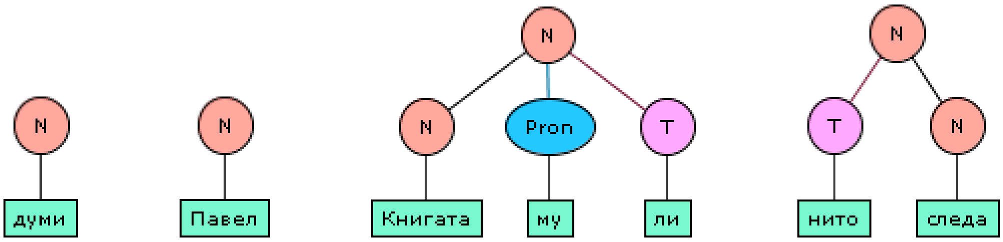
Gerund lexical element (*Gerund*) corresponds to a single gerund or to a gerund with clitics. The single word form is annotated with the lexical element *Gerund* if its morpho-syntactic tag maps the following pattern: "**V%%%g**" where **V** maps to the letter V, **g** maps to the letter g, **%** maps to any letter. In the graphical view we represent such word forms in the following way:



The gerund form shares a lot of features with participles. Having the same selectional potential as the verb lexeme, each gerund can have clitics (reflexive, not-reflexive — accusative and dative, interrogative, negative). There are no analytical gerund forms. The unexpressed subject is assumed to be coreferent with the subject of the clause that gerund or its maximal phrase modifies. Gerunds can be conjoined to form lexical coordination (*CoordP* element) which is also considered as a gerund. A gerund can be modified by the so-called emphasising words like 'само' (only) and the result is a gerund again.

3.4 Noun Lexical Element

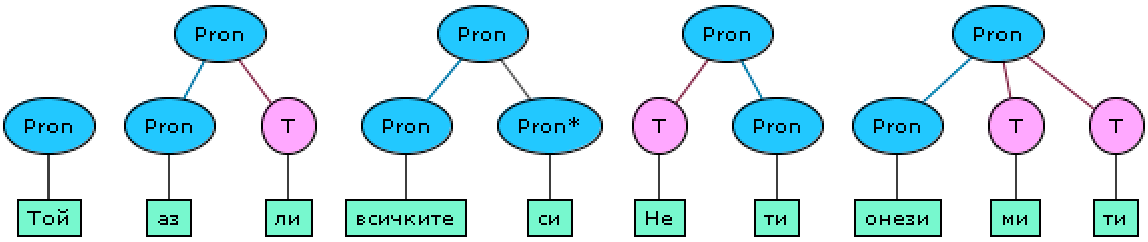
Noun lexical element (*N*) corresponds to a single noun or to an noun with clitics (possesive, interrogative, negative). The single word form is annotated with the lexical element *N* if its morpho-syntactic tag maps the following pattern: "**N#**" where **N** maps to the letter N, **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:



A noun can be a common noun, a proper noun — the first and the second example, respectively. It can have a possesive clitic and interrogative clitic as in the third example and it can have a negative clitic as in the fourth example. Nouns can be conjoined to form lexical coordination (*CoordP* element) which is also considered as a Noun. A noun can be modified by the so-called emphasising words like 'само' (only) and the result is a noun again. Each *N* element has a *sort* attribute which determines whether it denotes a Named Entity. In this case the value of the attribute has to be one of: *NE-Pers* for persons, *NE-Loc* for locations, *NE-Org* for organisations, *NE-Other* for other entities. In case of a common noun the value for the attribute by default is *common*.

3.5 Pronoun Lexical Element

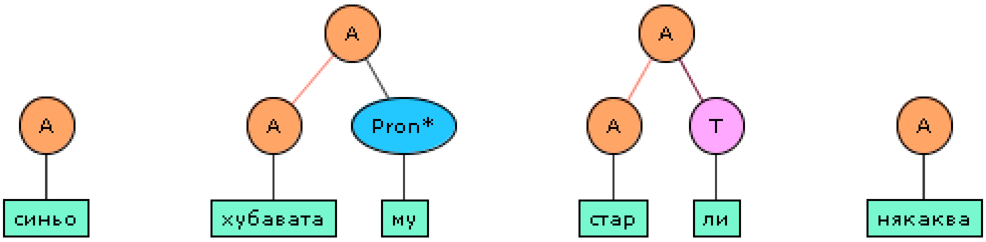
Pronoun lexical element (*Pron*) corresponds to a single pronoun or to a pronoun with clitics (possesive, interrogative, negative). The class of pronouns includes different kinds of words which we divide generally into three groups: proper pronouns, adjectival pronouns and adverbial pronouns. Here we present the proper pronouns. The other kinds of pronouns are presented below in the section on adjectives and the section on adverbs. The single word form is annotated with the lexical element *Pron* if its morpho-syntactic tag maps the following patterns: "**Pp#**", "**Pde#**", "**Pre#**", "**Pce#**", "**Pie#**", "**Pfe#**", "**Pne#**", "**Psot#**", "**Pszt#**", "**Psxt#**" where **P** maps to the letter P, **c** maps to the letter c, **d** maps to the letter d, **e** maps to the letter e, **f** maps to the letter f, **i** maps to the letter i, **l** maps to the letter l, **n** maps to the letter n, **o** maps to the letter o, **s** maps to the letter s, **r** maps to the letter r, **x** maps to the letter x, **z** maps to the letter z, and **#** maps to any sequence of letters. Thus a defining characteristic for a proper pronoun is that its referent type to be an entity (person, object) or to be a clitic. Sometimes such a pronoun can be used non-referentially and this is annotated via the attribute *ref* which has two possible values — *yes* (for refering pronouns) and *no* (for non-refering pronouns). The value *yes* is a by default value for the attribute. In the graphical view we represent such word forms in the following way:



The first example is the simplest case. The second one shows a pronoun with an interrogative clitic. The third example is a pronoun with a possesive clitic. The fourth example is a pronoun with a negative clitic. The last example is a special construction with two particles. Pronouns can be conjoined to form lexical coordination (*CoordP* element) which is also considered as a Pronoun. A pronoun can be modified by the so-called emphasising words like 'само' (only) and the result is a pronoun again.

3.6 Adjective Lexical Element

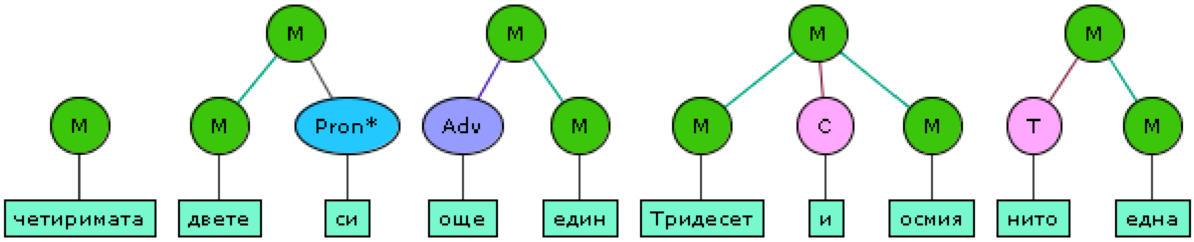
Adjective lexical element (*A*) corresponds to a single adjective, a single adjectival pronoun or an adjective with clitics (possesive, interrogative, negative). The single word form is annotated with the lexical element *A* if its morpho-syntactic tag maps the following patterns: "**A#**", "**Pda#**", "**Pra#**", "**Prp#**", "**Pca#**", "**Pcq#**", "**Pia#**", "**Pip#**", "**Pfa#**", "**Pfp#**", "**Pna#**", "**Pnp#**", "**Psol#**", "**Pszl#**", "**Psxl#**", where **A** maps to the letter A, **P** maps to the letter P, **a** maps to the letter a, **c** maps to the letter c, **d** maps to the letter d, **f** maps to the letter f, **i** maps to the letter i, **l** maps to the letter l, **n** maps to the letter n, **o** maps to the letter o, **p** maps to the letter p, **q** maps to the letter q, **r** maps to the letter r, **s** maps to the letter s, **x** maps to the letter x, **z** maps to the letter z, and **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:



The first example is a single adjective. The second example shows an adjective with a possesive clitic. The third example is an adjective with an interrogative clitic. The last example is an adjectival pronoun. Adjectives can be conjoined to form lexical coordination (*CoordP* element) which is also considered as an adjective. An adjective can be modified by the so-called emphasising words like 'само' (only) and the result is an adjective again.

3.7 Numeral Lexical Element

Numeral lexical element (*M*) corresponds to a single numeral, analytical numerals or a numeral with clitics (possesive, interrogative, negative). Analytical numerals consist of a numeral connected with the conjunction 'и' (and). Here is one example of such a numeral: тридесет и осмият (thirty eighth). The single word form is annotated with the lexical element *M* if its morpho-syntactic tag maps the following patterns: "**M#**", where **M** maps to the letter M, and **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:



The first example is the simples case of one numeral element only. The second example is a numeral with a possesive clitic. The third example is a numeral with an emphasising word. The fourth example is an analytical numeral. The last example is a numeral with a negative particle. Numerals can be conjoined to form lexical coordination (*CoordP* element) which is also considered as a numeral. A numeral can be modified by the so-called emphasising words like 'само' (only) and the result is a numeral again. Numerals are also presented in figures like: **30**, **3.14159**, **2001**, **1994-1996**, **1,72**, **30 000**, **18:23**; or figures with endings like: **7-и** (7th).

3.8 Family Name or Adjective, Derived from Family Names, Lexical Element

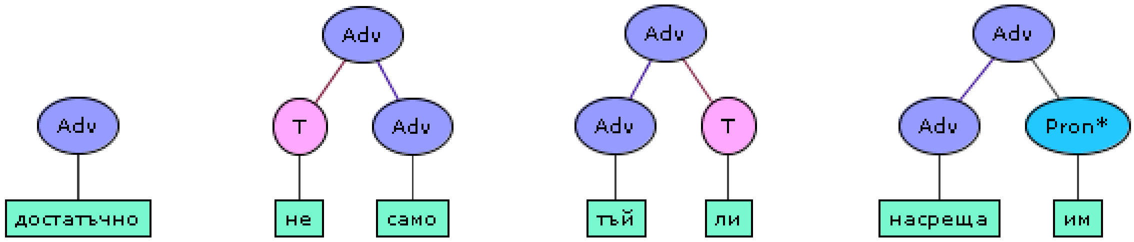
Family name or adjective, derived from family names, lexical element (*H*) corresponds to a single family name (or a derived adjective) or a family name (adjective) with clitics (possesive, interrogative, negative). Also here we include the Bulgarian second name which has the same properties as the family name. The single word form is annotated with the lexical element *H* if its morpho-syntactic tag maps the following patterns: "**H#**", where **H** maps to the letter H, and **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:

H
H
Златев
Костовият

The first example demonstrates an *H* element which can be a family name or an adjective and this depends on the usage. The second example shows an adjective from a family name. When an *H* element is used as a name this is stated via the attribute *sort* with a value *NE-Pers*. The name or the adjective can have an interrogative, possesive or negative clitic. The possesive and the negative clitics are more frequently when *H* element is used as adjective. *H* elements can be conjoined to form lexical coordination (*CoordP* element) which is also considered as an *H* element. An *H* element can be modified by the so-called emphasising words like 'само' (only) and the result is an *H* again.

3.9 Adverb Lexical Element

Adverb lexical element (*Adv*) corresponds to a single adverb, a single adverbial pronoun or an adverb with clitics (interrogative, negative, dative). The single word form is annotated with the lexical element *Adv* if its morpho-syntactic tag maps the following patterns: "**D#**", "**Pd%**", "**Pr%**", "**Pc%**", "**Pi%**", "**Pf%**", "**Pfq#**", "**Pfy#**", "**Pn%**", where **D** maps to the letter D, **P** maps to the letter P, **c** maps to the letter c, **d** maps to the letter d, **f** maps to the letter f, **i** maps to the letter i, **n** maps to the letter n, **q** maps to the letter q, **r** maps to the letter r, **y** maps to the letter y, and **#** maps to any sequence of letters. In the graphical view we represent such word forms in the following way:



The first example is a single adverb. The second example demonstrates an adverb with a negative particle. The third example shows an adverbial pronoun with an interrogative particle. The last example is an adverb with a dative clitic. Adverbs can be conjoined to form lexical coordination (*CoordP* element) which is also considered as an adverb.

There are some complex adverbs which are captured as a multi-word element (*mw*). Here is a list: като че ('kato che', as if), като че ли ('kato che li', as if), така нататък ('taka natatak', so on), все пак ('vse pak', all the same), все едно ('vse edno', it is all the same), може би ('mozhe bi', maybe), едва ли ('edva li', hardly), едва ли не ('edva li ne', hardly).

3.10 Conjunction Lexical Element

Conjunction lexical element (*C*) corresponds to a single conjunction. The single word form is annotated with the lexical element *C* if its morpho-syntactic tag maps the following patterns: "**C#**", where **C** maps to the letter C, and **#** maps to any sequence of letters. There are some complex adverbs which are captured as a multi-word element (*mw*). For a list of complex conjunctions see below. In the graphical view we represent such words forms in the following way:

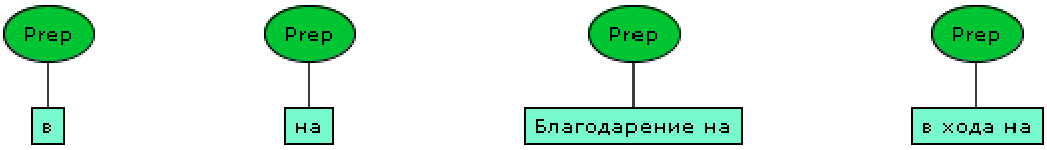
C
и
C
ако
C
за да
C
така и

The examples show two simple and two complex conjunctions. The first and the fourth example show coordinative ones, while the second and the third are subordinative.

Here is the list of the complex conjunctions in the treebank: ето защо ('eto zashto', that is why), за да ('za da', in order to), както и ('kakto i', as well as), камо ли ('kamo li', let alone), освен ако ('osven ako', unless), освен като ('osven kato', unless), освен че ('osven che', not only), само ако ('samo ako', only if), само че ('samo che', but), след като ('sled kato', after), сякаш че ('syakash che', as if), тъй като ('tyj kato', as), тъй че ('tyj che', so), така и ('taka i', thus), така че ('taka che', so). Note that some of the traditional complex conjunctions are represented as two word forms. Usually they consist of a preposition and a conjunction, or a preposition and auxiliary particle.

3.11 Preposition Lexical Element

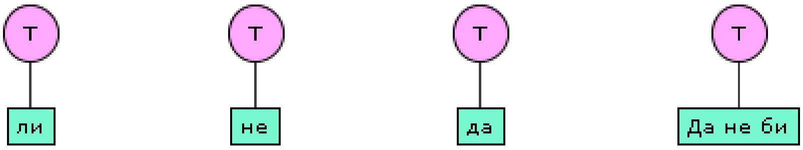
Preposition lexical element (*Prep*) corresponds to a single preposition. The single word form is annotated with the lexical element *Prep* if its morpho-syntactic tag maps the following patterns: "**R**", where **R** maps to the letter R. There are some complex prepositions which are captured as a multi-word element (*mw*). For a list of complex prepositions see below. In the graphical view we represent such word forms in the following way:



The first two examples are one token prepositions, the second two examples are complex prepositions. Here we give a list of some of the most frequent complex prepositions: във връзка с ('vav vrazka s', in connection with), в сравнение с ('v sravnenie s', in comparison with), в зависимост от ('v zavisimost ot', dependending on), в съгласие с ('v saglasie s', in accordance with), в хода на ('v hoda na' in the process of), с оглед на ('s ogled na', with a view to), по отношение на ('po otnoshenie na', with respect to), в съответствие с ('v saotvetstvie s', in conformity with), в близост до ('v blizost do', in the proximity of), за сметка на ('za smetka na', at the expence of), по време на ('po vreme na', at the time when), в полза на ('v polza na', in somebody's favour), в интерес на ('v interes na', 'in somebody's favour'), по случай ('po sluchaj', on the occasion of), в резултат на ('v rezultat na', as a result of), от страна на ('ot strana na', on the side of), по линия на ('po linija na', through, along), в памет на ('v pamet na', in memory of), благодарение на ('blagodarenie na', thanks to), etc.

3.12 Particle Lexical Element

Particle lexical element (*T*) corresponds to a single particle. The single word form is annotated with the lexical element *T* if its morpho-syntactic tag maps the following patterns: "**T#**", where **T** maps to the letter T, and **#** maps to any sequence of letters. There is a complex particle *да не би* (‘da ne bi’, in case) which is captured as a multi-word element (*mw*). In the graphical view we represent such word forms in the following way:



3.13 Interjection Lexical Element

Interjection lexical element (*I*) corresponds to a single interjection. The single word form is annotated with the lexical element *I* if its morpho-syntactic tag maps the following patterns: "**I**", where **I** maps to the letter I. In the graphical view we represent such word forms in the following way:

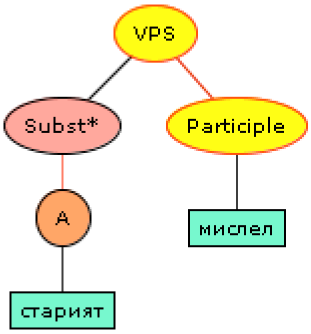
I
Ах
I
сбогом

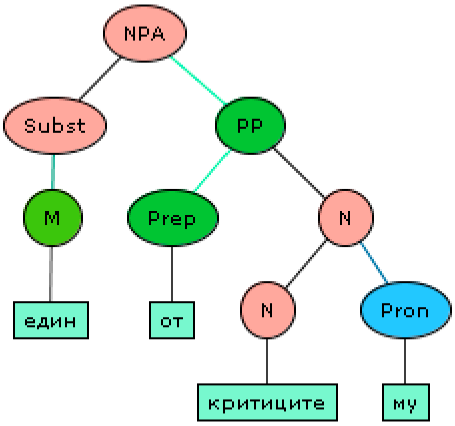
3.14 Type Shifting

Type-shifting covers phenomena of the following kind: when a word or a phrase of one category is used as a lexical item of another category. In our treebank type-shifting basically deals with three phenomena:

Substantivation

It is preferably lexicon-based and shifts the usual nominal dependands to heads. We substantivize the nominal elements like adjectives, numerals, participles, pronouns. A substativized element is annotated with the element *Subst*. It can take modifiers and form a noun phrase (see below). Here are some examples:

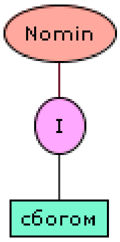


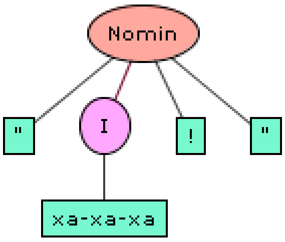


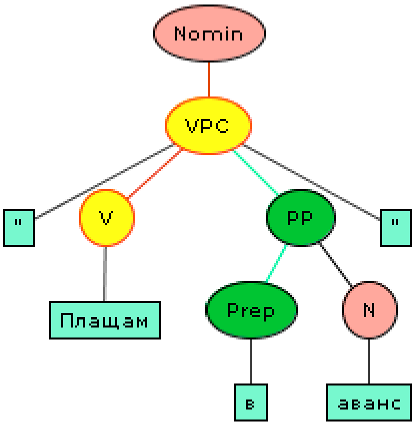
The first example shows a substantivation of an adjective to a noun. The second example demonstrates one of the typical phrases in the treebank which involves substantivation of a quantity word (a numeral) to a noun. This substantivation is governed by its own phrase: a quantity word is a head of NPA with a PP as a modifier, where the PP is formed with the preposition 'от' (from) and a NP (N in the example) denoting a set of entities. In this case the quantity word is substantivized to denote one or several of those entities. This type of co-referential relation is not explicitly encoded in the treebank, because it is easily recovered within the phrase itself.

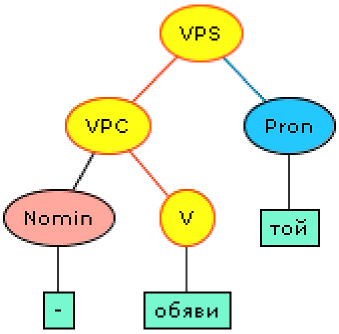
Nominalization

It is syntactically based and shifts predicates, intrejection and other non-nominals to nominals. Assuming that each predicate introduces a referent, very often we have other predicates that have this referent as an argument. We consider this process of converting of a predicate referent into a nominal referent a nominalization. Additionally, under this term we include converting of any phrase into a nominal (usually into a proper name). However, some of the cases of predicate nominalization are not annotated as such. Especially when a clause is referred to inside a sentense we use the same coreferential index for the clause and the coreferred nominal. But in cases when the clause is not in the same sentence we use explicitly nominalization. The last case is typical for newspaper texts in which there are a lot of examples of direct and reported speech. Usually the complement of the verb in the main clause consists of several sentences. We annotate this complement as nominalization. Nominalization is annotated with the *Nomin* element. Here are some examples:





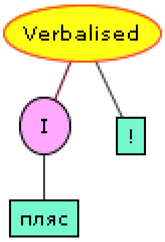


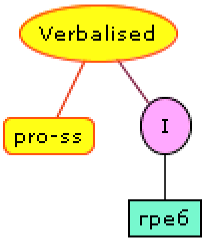


The first two examples demonstrate the nominalization of two interjections into nominals. The third example shows a nominalization of a title of a play ("I pay in advance"). The last example is the case of reported speech where the actual statement is outside the sentence. In this case the *Nomin* is either an empty element or it dominates a textual element that is connected to the statement — in the examle it is a dash.

Verbalization

This type-shifting appears when (usually) interjections, particles or adverbs express a predicate function. The most frequent case is of onomatopoeic interjections where the verbalised form substitutes the verb of the actual event. The verbalized form has the behaviour of a verb and can have a *pro-ss* element. The verbalization is annotated with the element *Verbalized.* Here are two examples:





The first example presents the basic case. The second exmple shows a verbalized element with a *pro-ss* element attached to it, which means that it is a part of a co-referential mechanism.

4 Verb Phrase

The verbal domain consists of the lexical element *V* and the phrasal elements: *VPC*(omplement), *VPS*(subject), *VPA*(djunct), *VPF*(iller). In this section we present the phrasal elements in the domain. Following the hierarchy of the realization of the dependents we have the following direction of realization: *lexical verb* -> *verbal head-complement phrase* -> *verbal head-subject phrase* -> *verbal head-adjunct phrase*. The lexical head of the phrasal element can be also a participle. It is done for consistency of the mapping between the lexical elements and the representation of Bulgarian morphology.

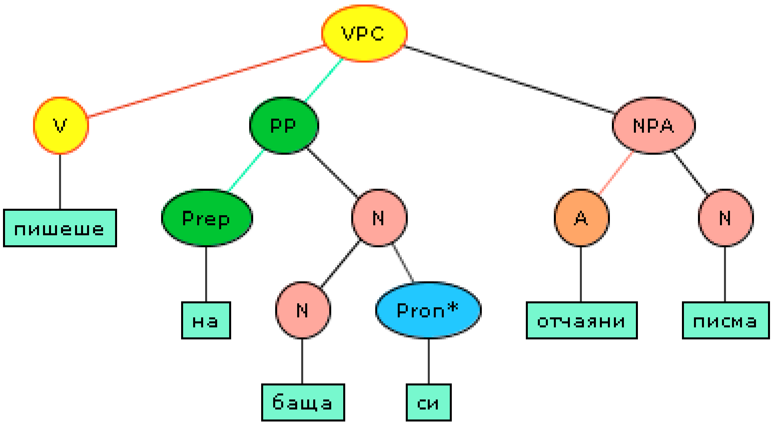
4.1 Head-Complement Verb Phrase (VPC)

VPC element corresponds to a verbal *head-complement* phrase. Each phrase of this type must have a head daughter and one or more complement daughters. All complements are realized at one step. Therefore, *VPC* can have only a lexical head and this includes the following elements: *V*, *Participle*, *V-elip*, *VD-elip*, *CoordP*, and *Verbalised*. The complements are of various kinds because we consider copula and copula like verbs as taking complements. Thus the possible complements include *nominals, adverbs, prepositional phrases, adverbials, clauses, participles, coordinated phrases*. A full list of these elements is presented below. The head daughter of a *VPC* can be easily determined because the set of the possible heads and the set of the possible complements share one common element — *Participle*. If a *VPC* consists of two participles we have to decide which one is the head. Even then the head can be determined with the help of more detailed morphological information. Here are some illustrations of *VPCs* consisting of two participles — детето било разглезено ('child-the was spoiled', They say that the child was spoiled); децата ставали разглезени ('children-the were becoming spoiled', The children were becoming spoiled)[1]. The head-dependent relation in both cases can be determined by the following assumption: the head participle is an active and the complement participle is a passive one. The head-dependent relation in a *VPC* which consists of coordination phrases can be determined by the propagation of the information from the arguments of the coordination phrase to the coordination phrase.

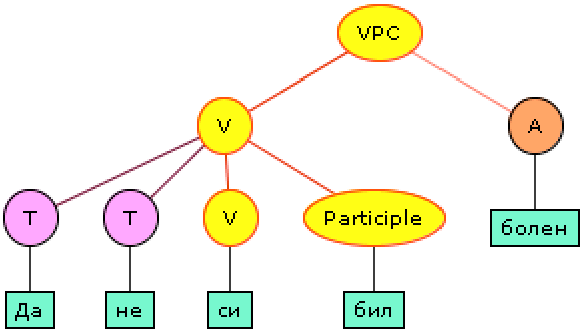
The complements can be realized in any linear order. Some complements can be extracted outside the *VPC* element and then they are represented as a *nid* element. If there are accusative and dative clitics within the verb

[1] Have in mind that in these cases the participles are ambiguous between passive constructions and predicative ones. Depending on the context, these illustrations can be interpreted as the one or the other. In our case we interpret them as predicative constructions. For a discussion see [Andreychin 1976].

head domain, the full-fledged complements have to agree in number, gender and person with them. The head and the complements can be discontinuous. They may be separated by *DiscA* elements (subjects or adjuncts of the same lexical head) or by *Pragmatic* elements (parenthetical expressions). The number and the kind of the complements depends on the properties of the lexical head. Here we first give two more general examples and then we present some special cases.



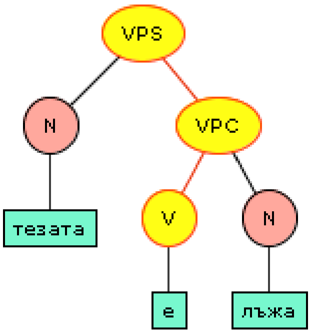
This is a verb head-complement phrase in which the lexical head (verb 'пишеше' (‘write’)) subcategorizes for two complements — a dative indirect object (indicated by a prepositional phrase) and an accusative direct object (indicated by a noun phrase).

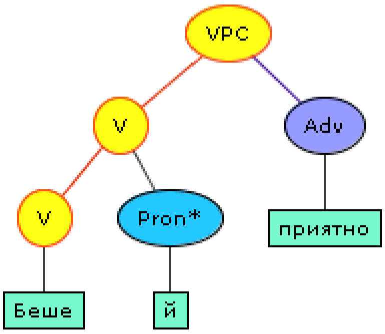


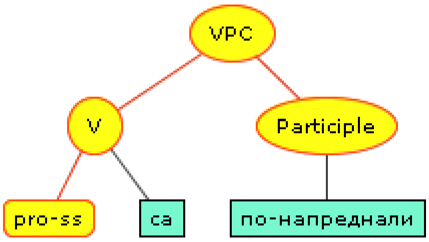
In this example the head is a verbal complex, which consists of the evidential perfect form of the copula ‘to be’ (си бил) and two clitics — the verb particle 'да' and the negative particle 'не'. The complement is an adjective.

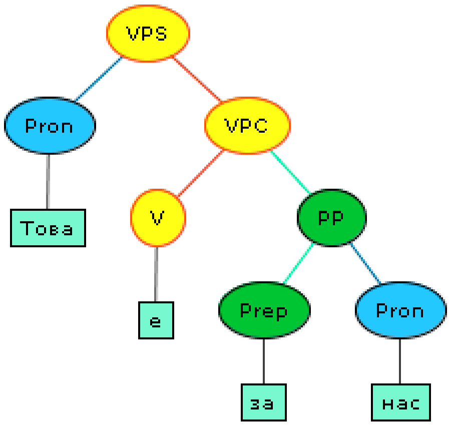
Copula *VPC*

We treat all predicative arguments (nominals, prepositional phrases, adverbials) of the copula as its complements. It is accepted by default that the complement of the copula co-refers with the subject in cases of predicative agreement. In many respects such a relation between the subject and the complement of the copula is similar to the relation between the subject and the agreeing adjuncts in secondary predication. Here we give some frequent examples:



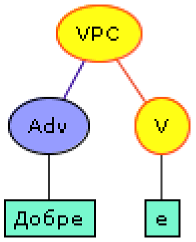


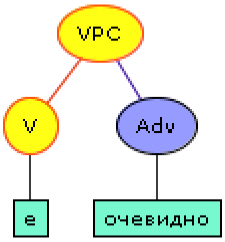




The examples above demonstrate various complement realizations of the copula. In the first tree, the complement is a noun. In the second one, the copula first forms a lexical sign with a dative clitic and then takes an adverb as its complement. The third example shows a participle in a predicative usage. Note that the head copula has a *pro-ss* element, which means that it is a part of a co-reference relation within the sentence. The last example presents a prepostional phrase as a complement of the copula.

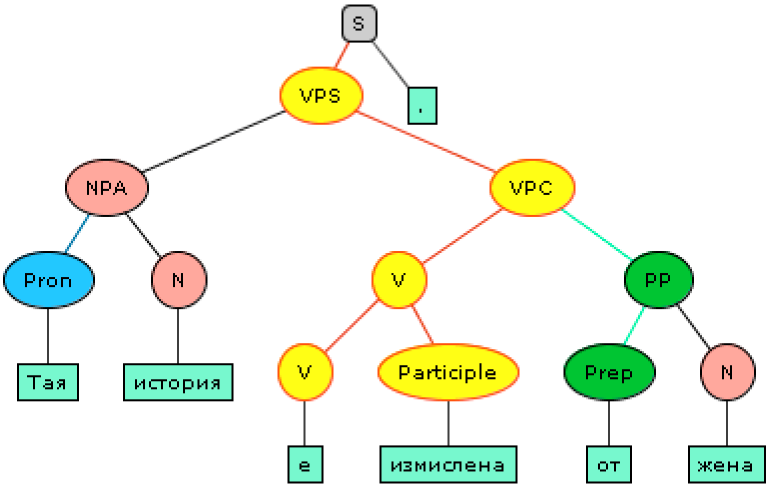
As *VPC*s we also treat impersonal constructions like: ясно е (it is clear), добре е (it is well), интересно е (it is interesting) etc. Note that there are cases when the copula is omitted: Естествено, че ще дойда = Естествено е, че ще дойда. Here are some examples:





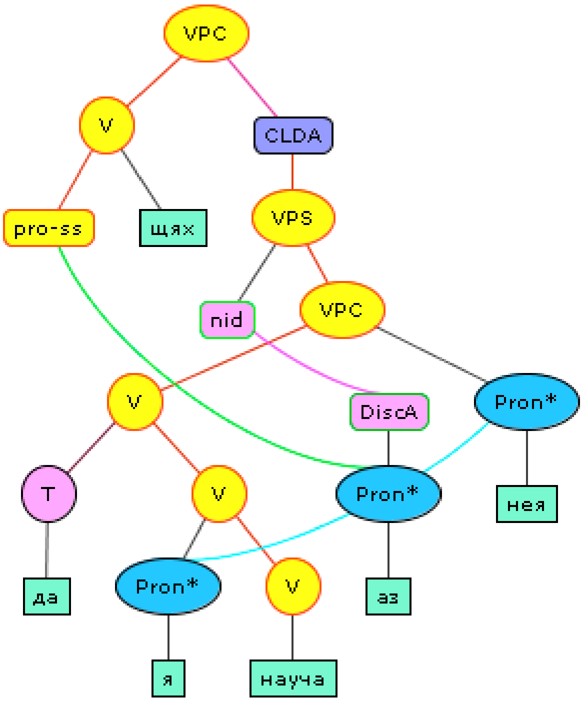
Passive constructions

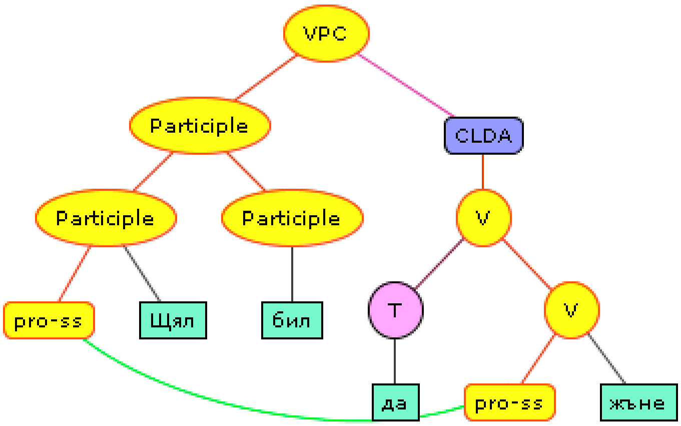
In the passive constructions we always treat the logical subject (prepositional phrase with the preposition 'от') as a syntactic complement. Thus we keep the surface realization of the verb head arguments. In the example below we present the passive variant of the active sentence: Жена измисли тая история (A woman thought of this story). The logical subject 'woman' is realized within a PP ('by a woman'), which is taken as a complement of the passive construction 'is thought of' ('This story was thought of by a woman').



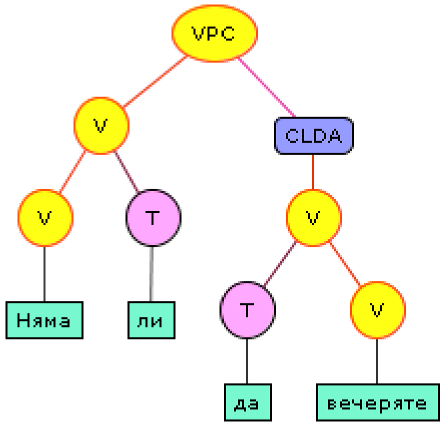
Analytical Verb Forms as VPC

As it was mentioned in the section on lexical verbs above, some traditional analytical word forms in the treebank are treated as head verbs subcategorizing for a clause. When the auxiliary verb in such a form heads a 'da' form we assume that the 'da' verb form constructs a clause and the auxiliary verb form selects this clause. Such cases are the analytical verb forms with the auxiliary verb 'ща' in its past tense — щях - (Future in the Past, Future Perfect in the Past, and their Intestimonial forms) and the negative forms for all future tenses 'няма, нямаше, нямало'. The motivation for this decision is that very often the subject and other dependents are realized locally within the 'da' form. In the first example below the subject and the complement are realized within the *CLDA* clause. According to the dependent realization principle, the complement is realized first. As the subject 'аз' ('I') separates the head verb and the complement, it is marked with the *DiscA* label and its constituent realization is pointed with the help of the *nid* element. Note that two co-reference relations are explicated as well: between the accusative clitic within the lexical verb sign and the full-fledged complement, and between the expressed subject of *CLDA* and the unexpressed one of the head verb щях (indicated as a *pro-ss* element).





In the second example we have an intestimonial form of the head verb. Note that the subjects of *CLDA* and the head verb are co-referent and for that reason both of them are represented as *pro-ss* elements.

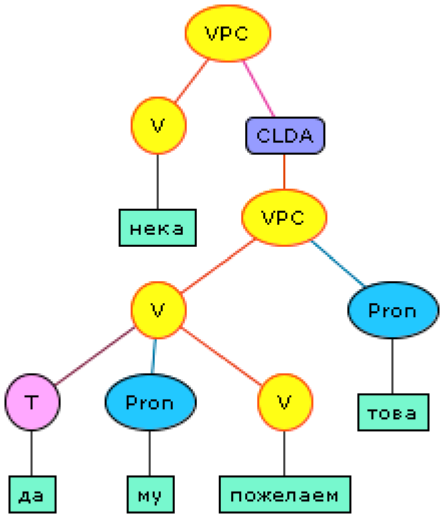


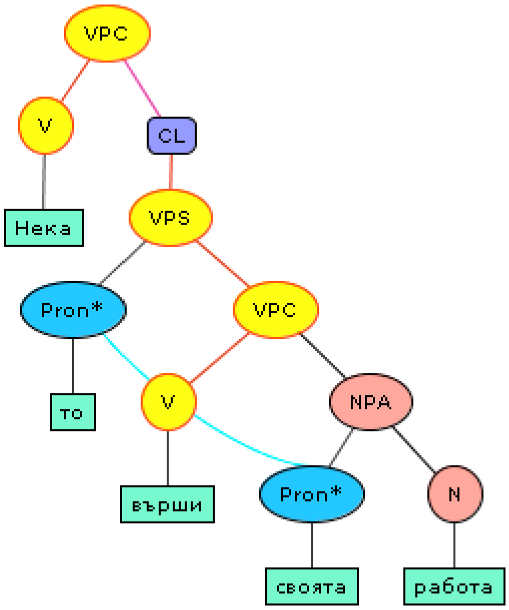
The last example shows a negative future form. It also illustrates the prosodic combination of an interrogative particle and the negative verb under the lexical *V* element. The negative verb няма is impersonal and for that reason it cannot structure-share a subject with the 'da' clause.

Clausal Complements

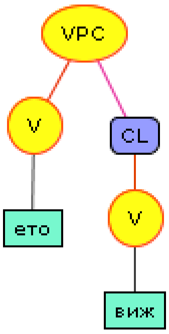
The verbs can take various clausal complements (*CLDA*, *CLQ*, *CLR*, and *CL*). It is in cases when the complement slot of a verb is occupied by a clause. The role of each clause as a complement is discussed below in the section on clauses. This is why here we present two special cases, in which the head is either a verbal particle, or an idiosyncratic one.

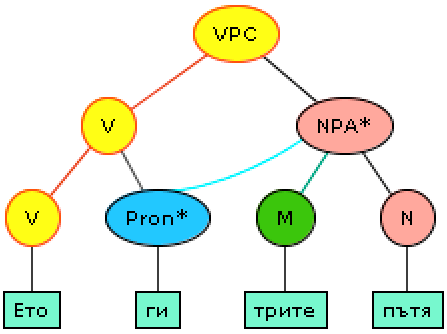
Verbal particles like 'нека' (‘neka’, let's) and 'ето' (‘eto’, here it is) take either *CLDA* or *CL* clauses as complements (‘eto’ can also take other types of complements). We annotate these kinds of particles with a *V* lexical element. Here are some examples:





In the first example the verbal particle 'нека' takes a *CLDA* as a complement. In the second example the verbal particle 'нека' takes a bare clause, i.e. without a 'da'-form. Note that in this case the *CL* marking is conventional and is used for consistency with the other analyses. In the third example (on the next page) the verbal particle 'ето' takes a conventional *CL* as a complement. In the fourth example (on the next page) 'ето' assigns an accusative case to its nominal complement.

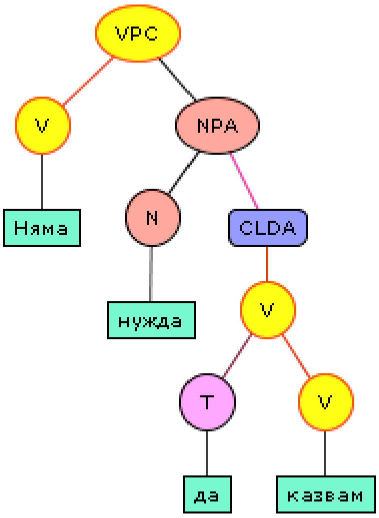




The verb 'стига' (‘stiga’, enough) takes only clauses of the general type *CL*. Note that when it is combined with *CLDA*, then it equals the subordinator 'ако' (‘ako’, if) and therefore heads a conditional adjunct clause.

Some Special Cases

Here we have in mind constructions with supportive and light verbs. The supportive verbs 'имам' ('imam', have) and 'нямам' ('nyamam', not have) form a *VPC* construction of special kind. Their first complement takes the second one as its modifier and just then is realized by the verb. Here is an example: the noun ‘need’ takes a *CLDA* as an adjunct and forms an *NPA*. Then the *NPA* becomes a complement of the verb.



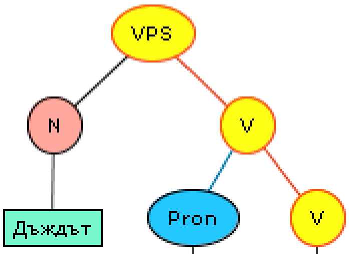
There are constructions with the so-called 'light' verbs. For example, държа реч пред някого (to make a speech to someone), държа врага под око (to keep the enemy under one's eye), хвърлям поглед на нещо (to have a look at something) etc. Note that such constructions exhibit different degrees of idiomaticity, therefore there are different relations between the verb and one or both complements, and between the complements themselves. Because we lack an exhaustive typology of these cases, we decided that when the first complement is not closely related to the second one and cannot take it as its own dependent, the head verb takes two complements.

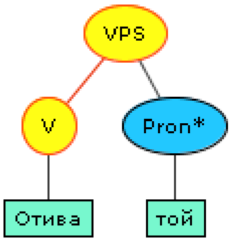
4.2 Head-Subject Verb Phrase (VPS)

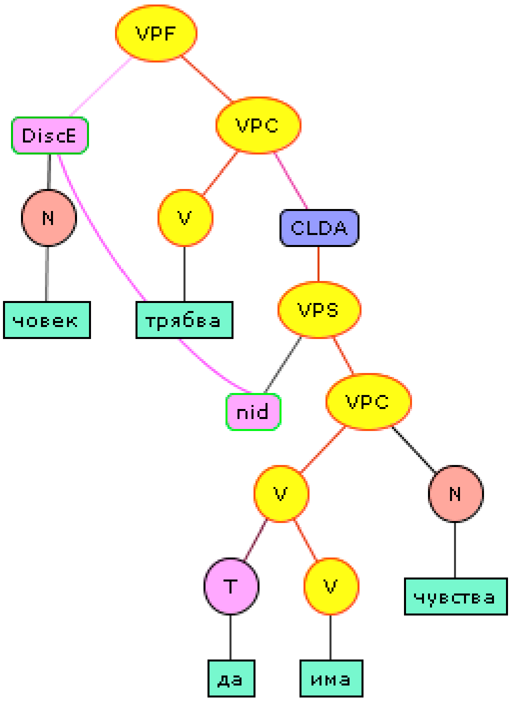
VPS element corresponds to a verbal *head-subject* phrase. Each phrase of this type must have a head daughter and a subject daughter. *VPS* can have a lexical head if there are no complements. The head variety includes the following elements: *VPC*, *V*, *Participle*, *V-elip*, *VD-elip*, *CoordP*, and *Verbalised* (i.e., all possible heads for a *VPC* element or the *VPC* itself). The possible subject daughters include *nominals* (plus the substantivized and nominalized elements), *clauses, prepositional phrases, adverbials, coordinated phrases*. We consider some types of *prepositional phrases* and *adverbials* as subjects, i.e. in cases when they express a quantity or proximity of objects or material. Then we accept that they have a nominal referent. As it can be seen, the head of a *VPS* is easily determined because the possible heads and the possible subjects do not overlap. The head in the case of a

VPS consisting of coordination phrases can be determined by propagation of the information from the arguments of the coordination phrases to the coordination phrases.

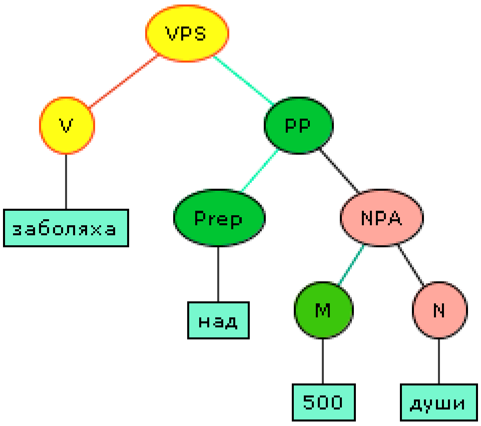
Generally, the subject can be realized at any place within the sentence. Of course, there are some local restrictions like the following: the subject cannot be placed immediately after the ‘da’-form in 'da'-constructions (*да Петър дойде). Some subject can be extracted outside the *VPS* element and then its actual constituent place is represented as a *nid* element. The subject usually agrees with the lexical verb head in number, gender and person[2]. The head and the subject might be discontinuous. They can be separated by impersonal verbs or *DiscA*, *DiscE* elements (adjuncts of the same lexical head, constituent from the complement of the same lexical head) or by *Pragmatic* elements (parenthetical expressions). The syntactic type of the subject depends on the lexical head. Here we give examples of each major subject type.



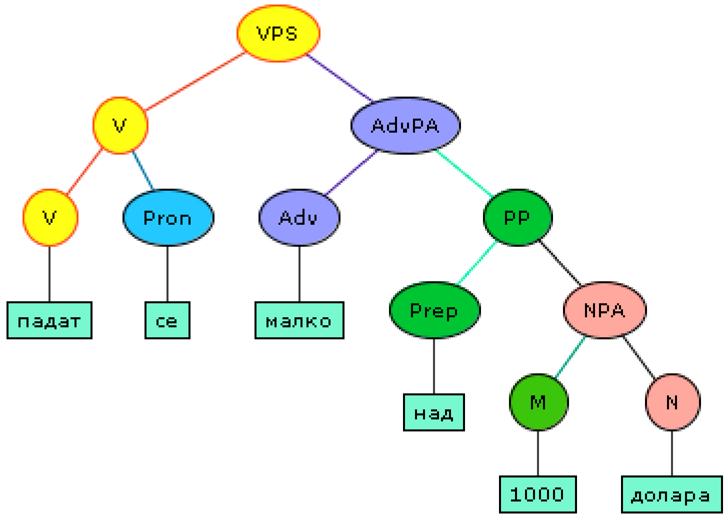




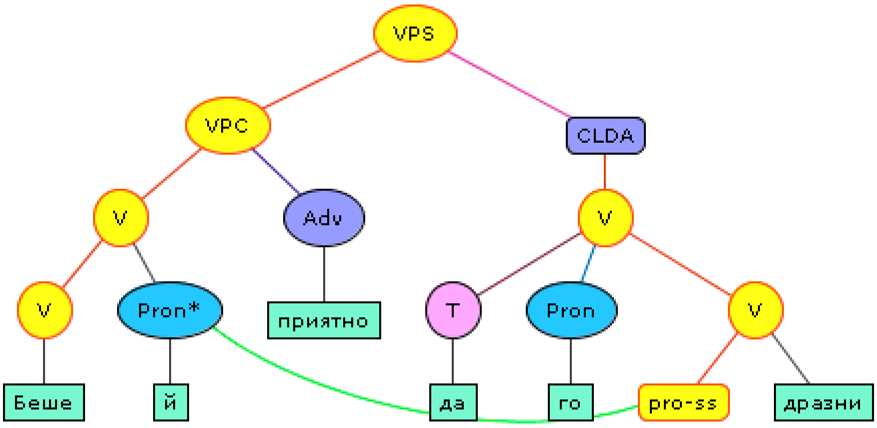
In the examples above the subjects are nominals, a noun and a pronoun respectively. The second example shows a reverse word order, namely a verb-subject one. The third example demonstrates an extracted subject in a construction with impersonal verb selecting a 'da'-clause. In the following two examples the subject is expressed by an adverbial and by a prepositional phrase. If we remove the prepositions and the adverb, the nominal groups show agreement with the head verb.



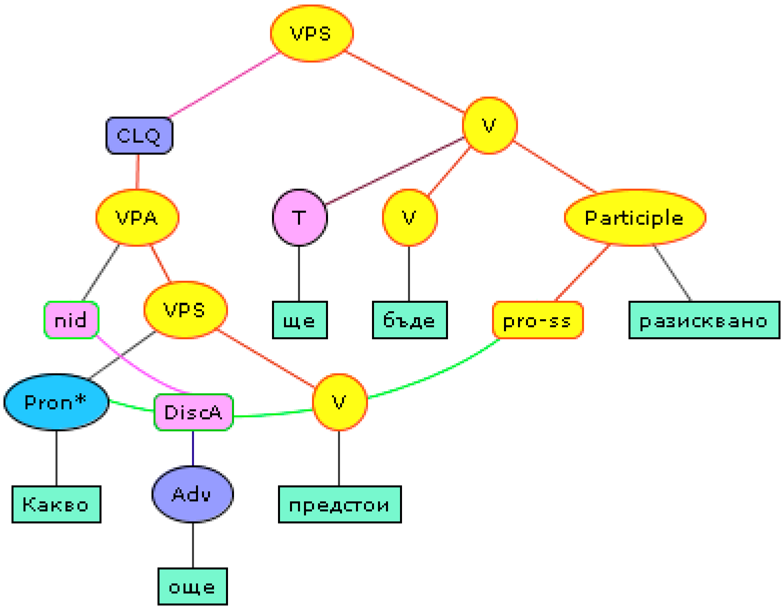
[2] There are deviations from this rule. For a discussion see in [Bulgarian Academy Grammar 1983].

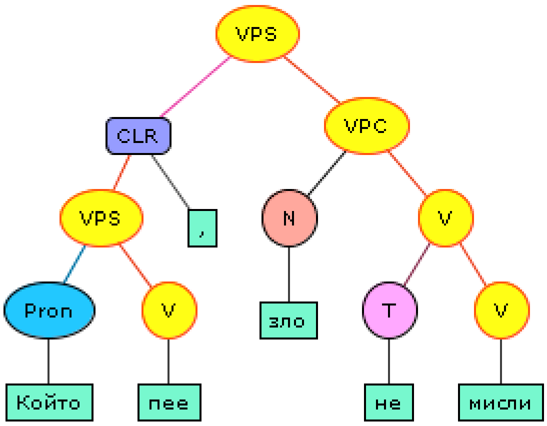


When the subject is a clause (*CL*, *CLDA*, *CLCHE*, *CLQ*, *CLR* see below for concrete examples) there are generally two cases. The first is that the event referent introduced by the clause is the subject referent. The second case is when the subject of the subordinated clause is co-referent with the subject of the main clause. The first case is always applicable to *CL*, *CLDA*, *CLCHE* clauses. The second case always holds for the relative clauses (*CLR*). Interrogative clauses (*CLQ*) can be of both cases. For that reason, if there are co-references between the subjects of the *CLQ* clause and the main sentence, we always explicate such coreferences. We never explicate the coreference if the subject clause is of type *CLR*, because it is always infereable. To be more precise, in these cases the main clause does not have its own subject, but the relative or the interrogative clauses modify the appropriate element on the **ARG-ST** list. We represent this fact by a co-referential relation between the subject of the subordinated clause and the *pro-ss* element of the main clause. Below we present three examples:



The first example is of the first kind. The referent of the whole *CLDA* clause is the subject referent of the main clause. Note that the unexpressed subject of the *CLDA* is structure-shared with the dative clitic in the main sentence.The next two examples demonstrate the co-referential relations between the subject in the subordinated clause and the *pro-ss* element in the main clause. In the first example, we have a *CLQ* clause which shares the referent of its subject with the *pro-ss* element in the main clause. This relation is explicated, because it is not obligatory in principle. In the second example we have a *CLR* clause as a subject. In this case it is obligatory for the subject's referent of the *CLR* clause to coincide with the corresponding element of the **ARG-ST** list of the verb of the main clause. This relation is not shown because it can be easily reconstructed.

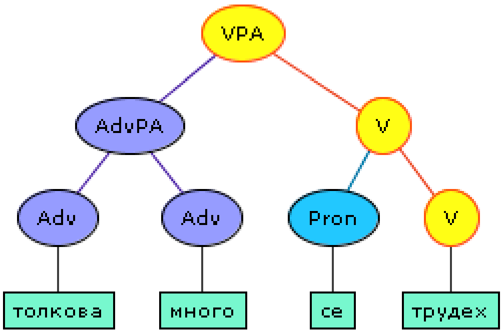


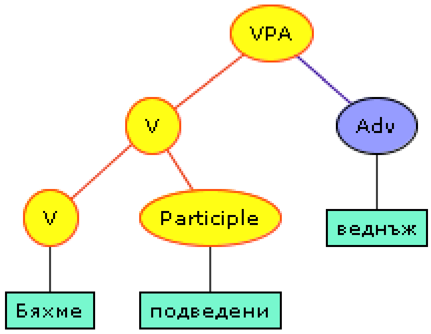


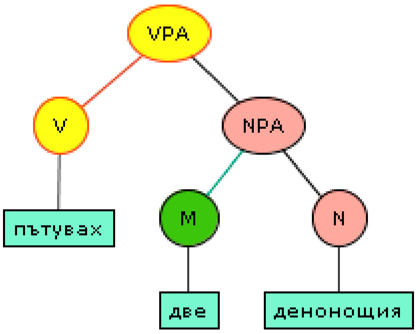
4.3 Head-Adjunct Verb Phrase (VPA)

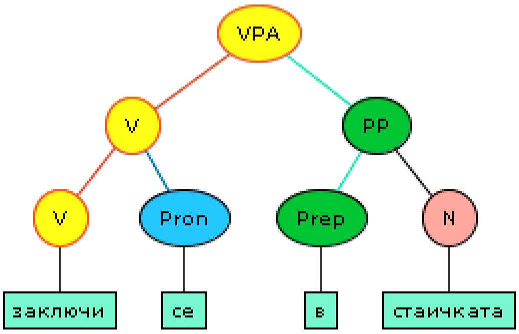
VPA element corresponds to a verbal *head-adjunct* phrase. Each phrase of this type must have a head daughter and an adjunct daughter. *VPA* can have a lexical head if there are no complements and a subject, a *VPC* head, *VPS* head or VPA head. The list of possible heads includes the following elements: *VPA*, *VPS*, *VPC*, *V*, *Participle*, *V-elip*, *VD-elip*, *CoordP*, and *Verbalised* (all posible heads for a *VPC* element, *VPS* and *VPA*). The possible adjuncts include nominals (plus substativized and nominalized elements), adjectives and adjective phrases, clauses, prepositional phrases, adverbials, participles, particles, coordinated phrases. Because the list of the possible heads of a *VPA* and the list of the possible adjuncts have a common element (*Participle*), again there will be a problem to determine the head of *VPA* when it consists of two participles of the same kind. It is possible, but very rare, to have one participle as a lexical head and a second participle as a secondary predication — стоял онемял ('stoyal onemyal', He was standing speechless). In our view, the sentenses of this kind have only one word order possibility — first comes the lexical head and then the adjunct. Thus we can stipulate that in such cases always the first participle is the head. The head in the case of a *VPA* consisting of coordination phrases can be determined by propagation of the information from the arguments of the coordination phrases to the coordination phrases themselves.

The adjunct can be realized in any order (except in cases of a *VPA* consisting of two participles). It can be extracted outside the *VPA* element or it can be realized among the other dependents of the same lexical head and then it is represented as a *nid* element. In case of secondary predication the adjunct agrees with the subject or one of the complements of the lexical verb head in number, gender and person. The head and the adjunct can be discontinuous. They can be separated by *DiscE* elements (a constituent from the subject or a complement of the lexical head) or by *Pragmatic* elements (parenthetical expressions). Here we give some examples of adjuncts.



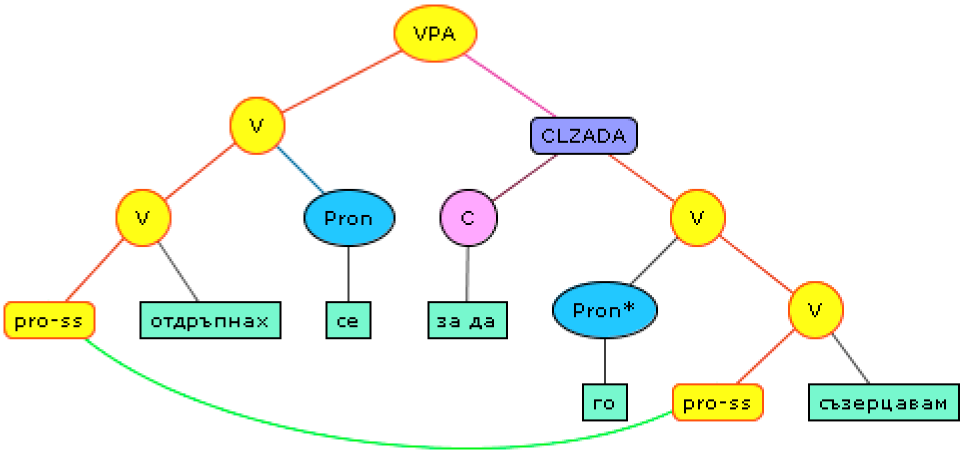


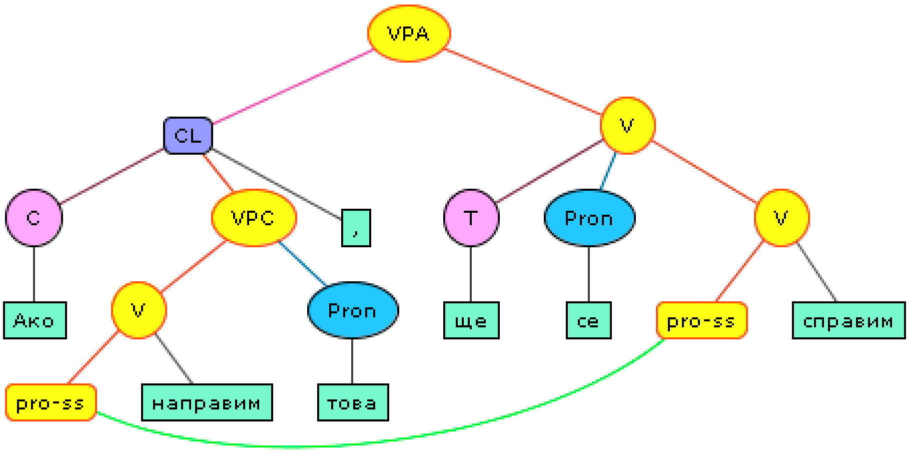




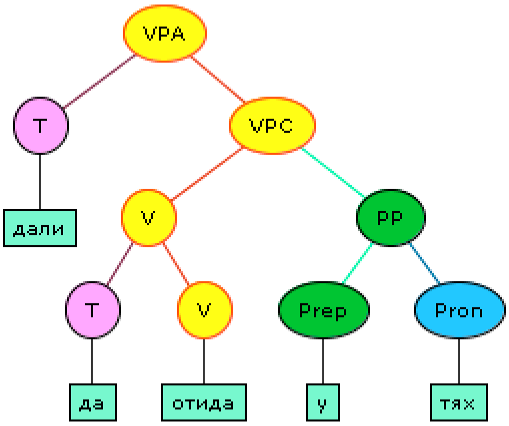
The first two examples above demonstrate adverbial adjuncts. The third example shows a nominal adjunct denoting a time period. The fourth example below shows an adjunct for place expressed by a prepositional phrase.

The clausal adjuncts can be of different nature, but the most frequent ones are for purpose and for condition. Here the first clause is for purpose, the so called *CLZADA* clause, introduced by the complex conjunction 'за да' ('za da', in order to). The second example is with a conditional clause, annotated as a general clause *CL*. In cases like these when there are co-referential relations between the unexpressed subjects of the clause and the matrix sentence we assign two *pro-ss* elements and link them.





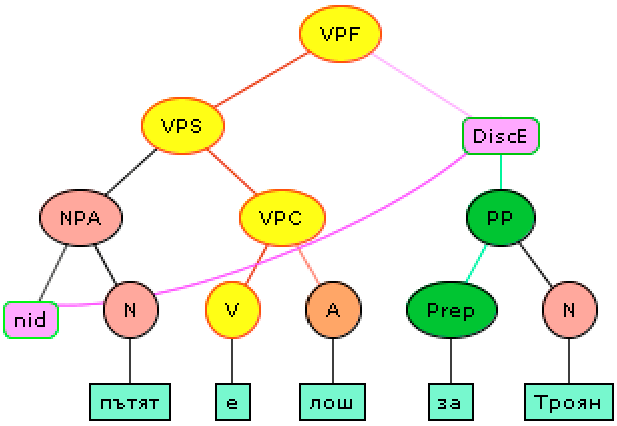
A specific type of *VPA* are those, projected by the interrogative particles: 'нали' ('nali', isn't it?), 'дали' ('dali', whether), 'нима' ('nima', really?) and the modal particle 'да не би' ('da ne bi', in case). Some of them also play a pragmatic role and then they are annotated as *Pragmatic* element.



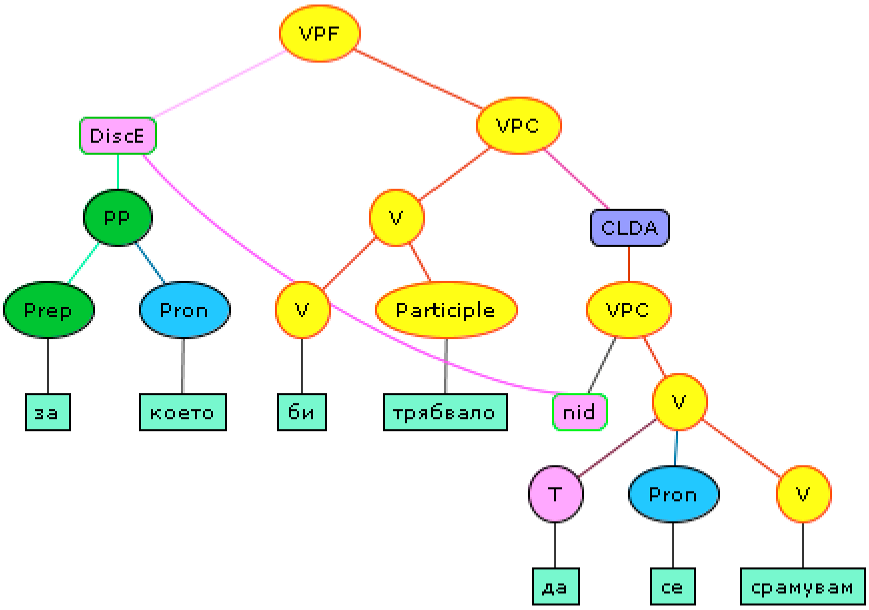
4.4 Head-Filler Verb Phrase (VPF)

VPF element — corresponds to a verbal *head-filler* phrase with some changes. Each phrase of this type must have a head daughter and one or more filler daughters. *VPF* can have a lexical head (if there are no complements, adjuncts and a subject), a *VPC* head, *VPS* head or *VPA* head. The list of possible heads includes the following elements: *VPA*, *VPS*, *VPC*, *V*, *Participle*, *CoordP*. The possible fillers are *DiscE* elements. Note that the *VPF* element does not correspond completely to the verbal head-filler phrase. First, when possible, the fillers are realized at once. Another difference is that when the filler is among the other dependents of the lexical head, we do not introduce a *VPF* element, but attach the filler (*DiscE* element) to the immediate dominance node. See below for examples. These differences between the HPSG head-filler phrase and the *VPF* are due to the following design principle: each piece of information that can be easily inferred from the represented information is not presented.

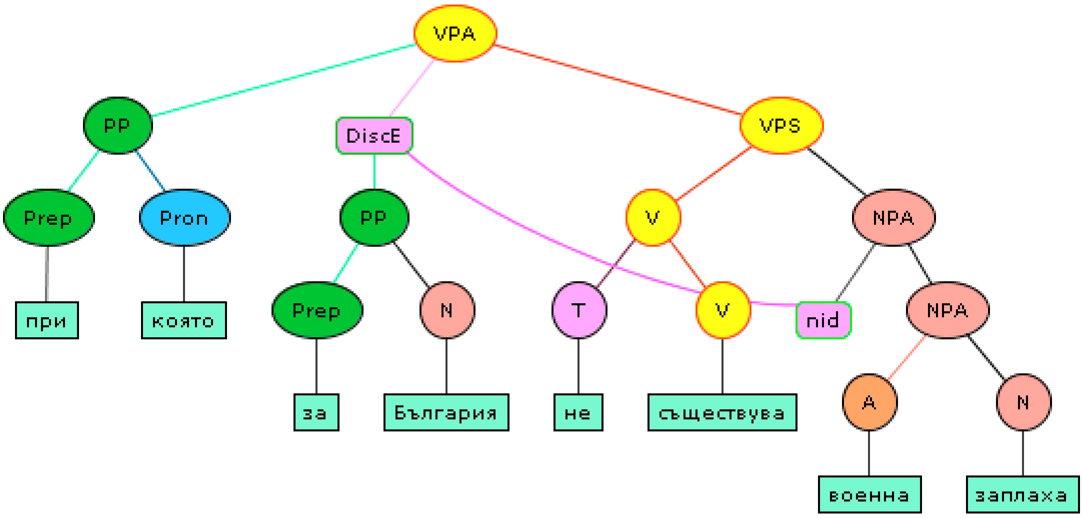
The fillers can be realized in any order. The head and the fillers can be discontinuous. They can be separated by *Pragmatic* elements (parenthetical expressions). Here we give some examples of *VPF* elements:



In this example we have an extraction of a *PP* modifier from a noun phrase outside the *VPS* element having the noun phrase as a subject. We consider such an extraction as a traceless analysis and thus, the *nid* element which marks-up the constituent from where the material was extracted, is always the first daughter. Recall that the *nid* element does not indicate a word order position, but just the syntactic one. The first of the next two examples is typical for relative clauses with an impersonal modal verb which selects for a clause. In this case the relative constituent (the constituent with a lexical head which is a relative pronoun) has to be first in the word order. Thus very often it is extracted. In the example the complement of the subordinated clause (*CLDA*) is realized on the level of the modal verb. Any dependent (complement, subject, adjunct) can be extracted from the subordinated clause in this way. The grammatical role depends on the relative pronoun.



The example on the next page demonstrates a *DiscE* element without a *VPF* element. In this case we have an extracted element, but among the dependents of the same verbal lexical head. In the example there is an extracted element from the subject, but the extracted element is realized between the *VPS* element and the adjunct. Thus *DiscE* is attached to the *VPA* element.



4.5 Clauses: Saturated Verb Phrase (CL)

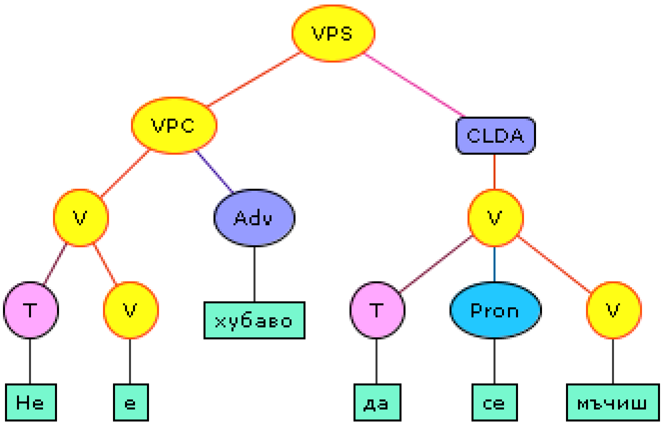
A verb phrase with realised dependents of the lexical head is called a saturated verb phrase — clause. A saturated verb phrase can be a complement, subject or adjunct of other phrases. In order to annotate the saturation we use a set of elements: *CL*, *CLDA*, *CLCHE*, *CLZADA*, *CLQ*, *and CLR*. The clauses are divided into few types according to the introducing element, not the function, although some of the types have a special behaviour. Note that the division is restricted to the following main types: *CLDA* — a clause which lexical head is 'da' form; *CLCHE* — a clause introduced by the conjunction 'че' ('che', that); *CLZADA* — a clause introduced by the conjunction 'за да' ('za da', in order to); *CLQ* — a clause introduced by a question word, *CLR* — a clause introduced by a relative pronoun. All the other clauses are marked with the general tag *CL*.

4.5.1 CLDA

CLDA is a clause, which is introduced by the verbal form 'da'. It can represent different dependent roles with respect to its head. For example:

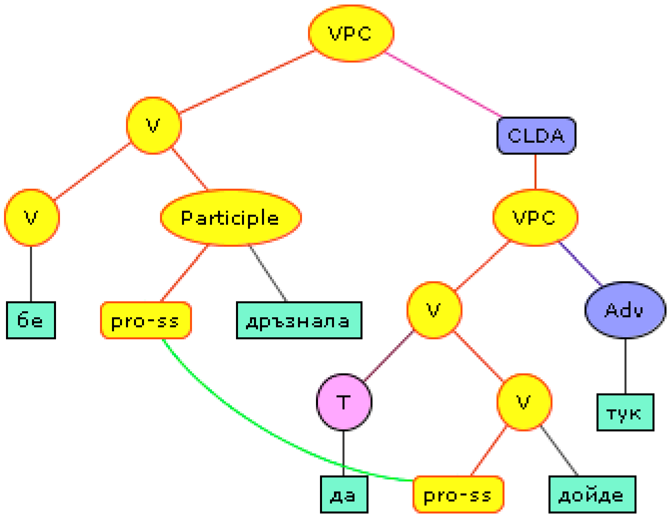
Subject

In the example below the VPC phrase takes a CLDA as its subject.



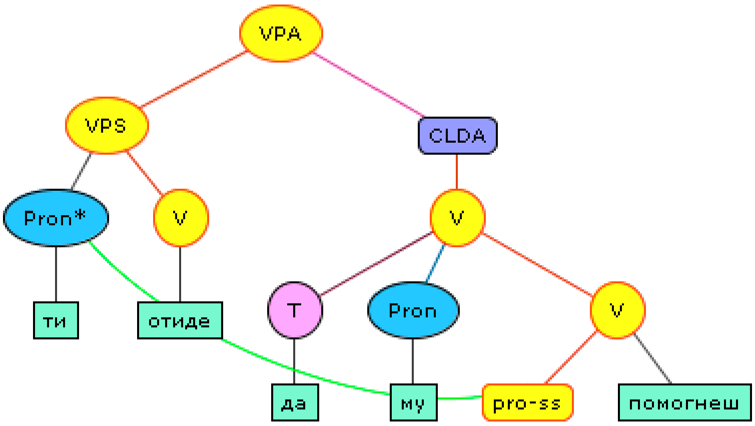
Complement

Apart from the treatment of the special analytical verbal forms discussed above there are other verbs that select for *CLDA* clauses. One interesting case are the control verbs. Here is one example of such a construction. We assume that the complement clause is saturated and thus it has an explicit or implicit subject which is co-referent with the subject of the main verb or one of its complements. In the example below the control is expressed by the co-referential relation between the two *pro-ss* elements.



Adjunct

In this position *CLDA* plays the role of an adjunct for purpose (in parallel to *CLZADA*). Here is an example, in which the subject of the main sentence is co-referred with the dative clitic in the subordinate clause:

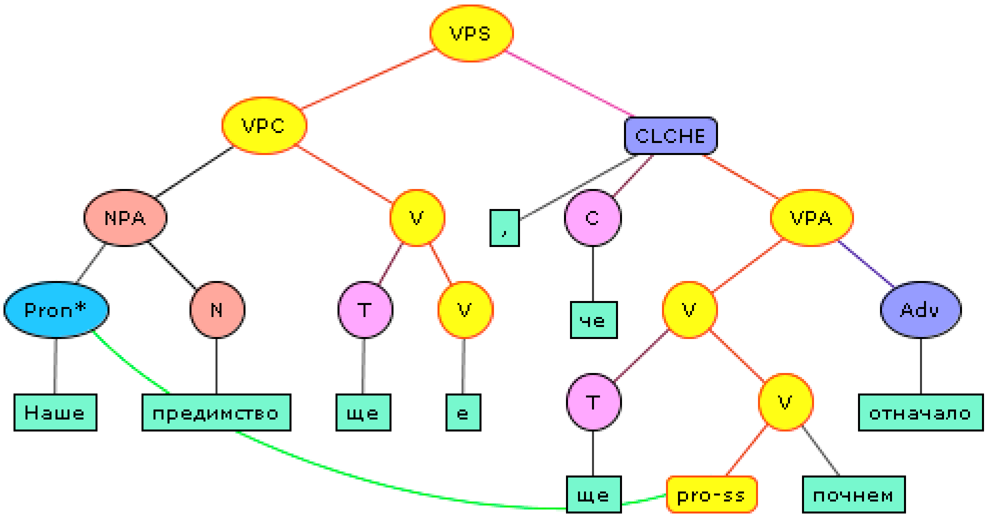


Note that when the da-construction is not a subordinate clause, then it is not marked as *CLDA*: Да тръгваме! (Let us go!); Когото го мързи, да не става! (Who is lazy, let him not stand!). Here the appropriate information comes from the illocutionary status of the sentence - imperative, optative.

4.5.2 CLCHE

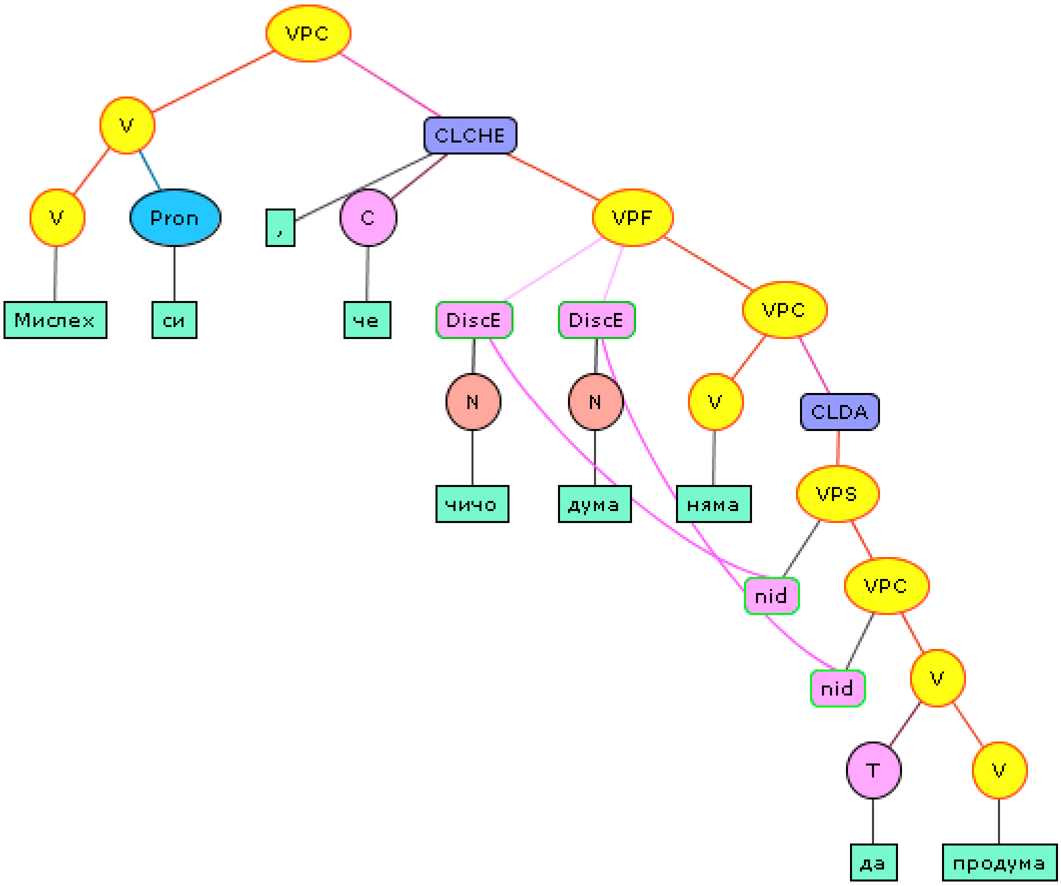
CLCHE is a clause, which is introduced by the conjunction *че* ('che', that). It can represent different dependent roles with respect to its head. Traditionally it is accepted that this conjunction is a complementizer, but the clauses introduced by it can play different roles and we do not annotate them in a special way. Here are some examples:

Subject



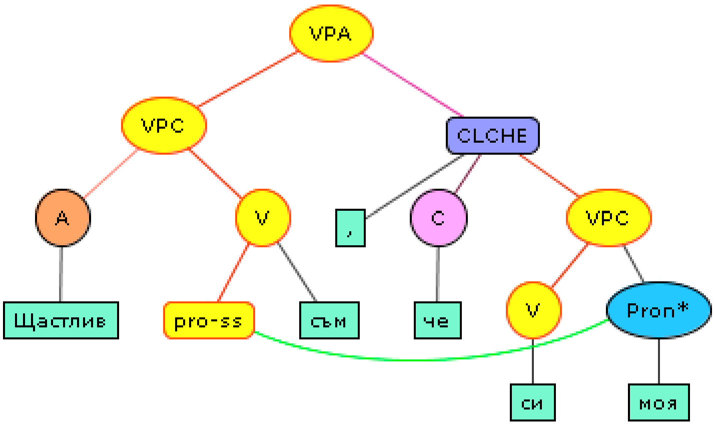
Complement

Some verbs select for such kind of clauses. A large group of verbs selecting for these clauses are the verbs of reporting. Here is one example, which also shows two extractions from one clause.



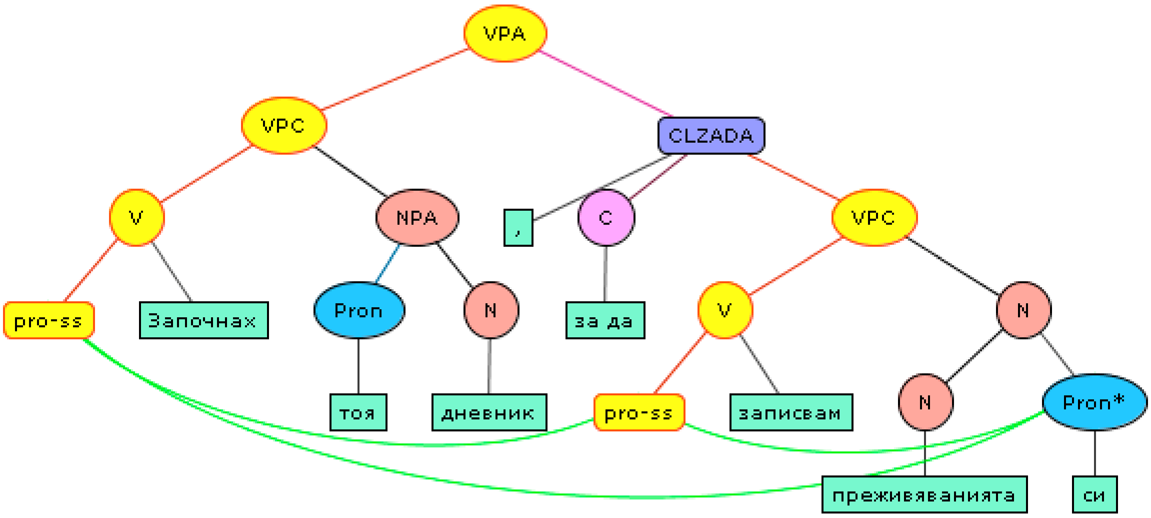
Adjunct

In the example below че introduces a clause of reason.



4.5.3 CLZADA

CLZADA is a clause, which is introduced by the conjunction 'за да' ('za da', in order to) and expresses the purpose of a certain action. It represents an adjunct dependant role with respect to its head. In the example below there is a triple co-referent relation between the unexpressed subjects of the main sentence and the subordinate clause, and the reflexive possessive clitic in he subordinate one:

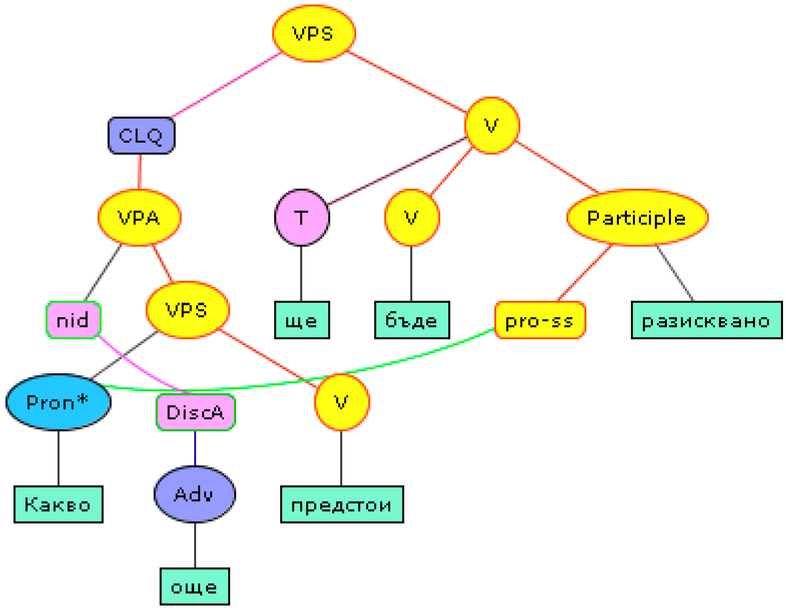


4.5.4 CLQ

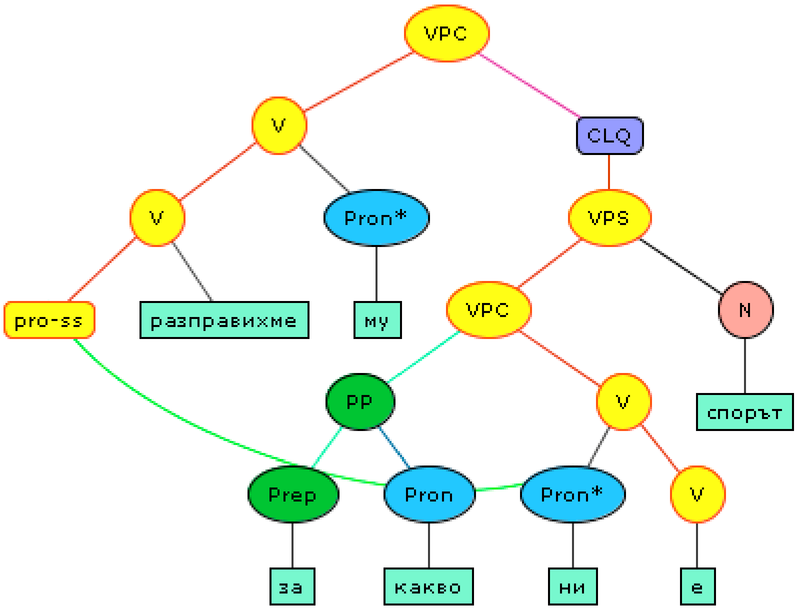
CLQ is a clause, which is introduced by an interrogative pronoun or particle. It can represent different dependent roles with respect to its head. For example:

Subject

In the example there is a co-referential relation between the subject of the *CLQ* clause and the unexpressed subject of the main verb.

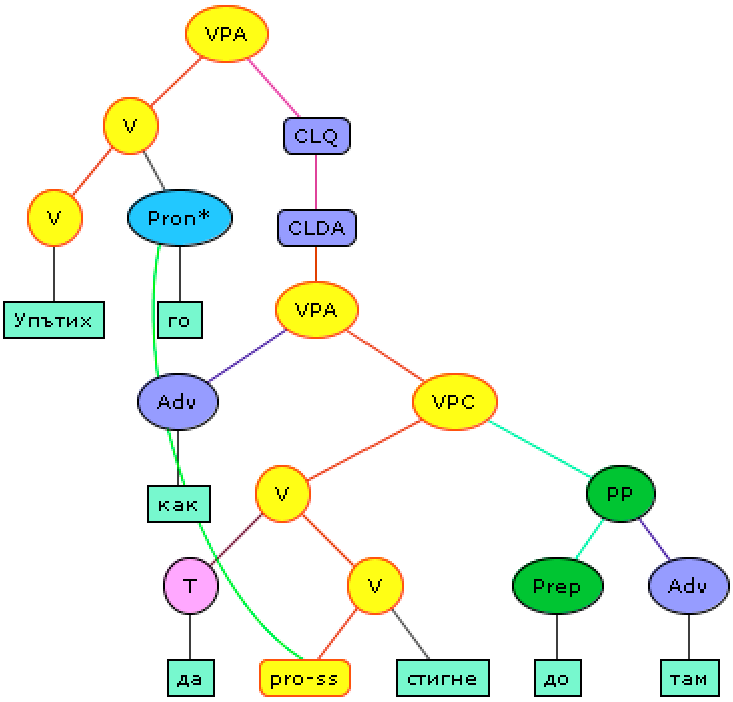


Complement



Adjunct

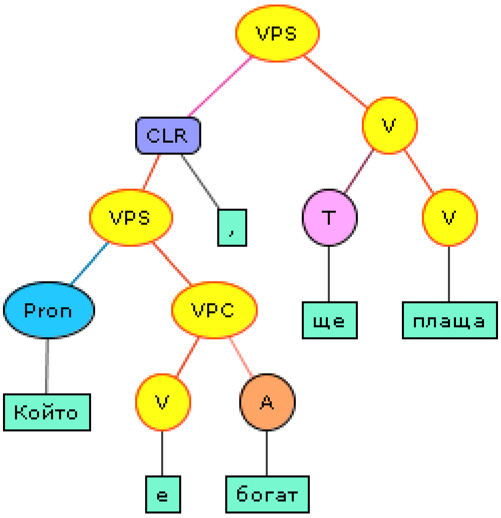
Note that in the example below the subordinate clause is double-marked as *CLDA* and *CLQ*.



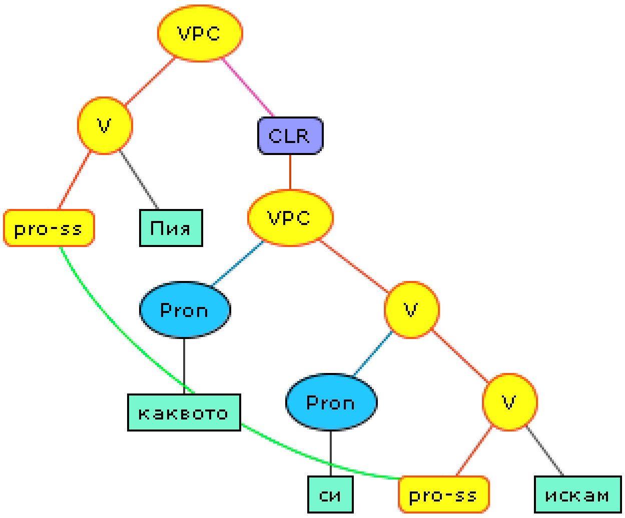
4.5.5 CLR

CLR is a clause, which is introduced by a relative pronoun and surrounded by commas. Relative clauses can be non-free and free. They can represent different dependent roles with respect to its head. For example:

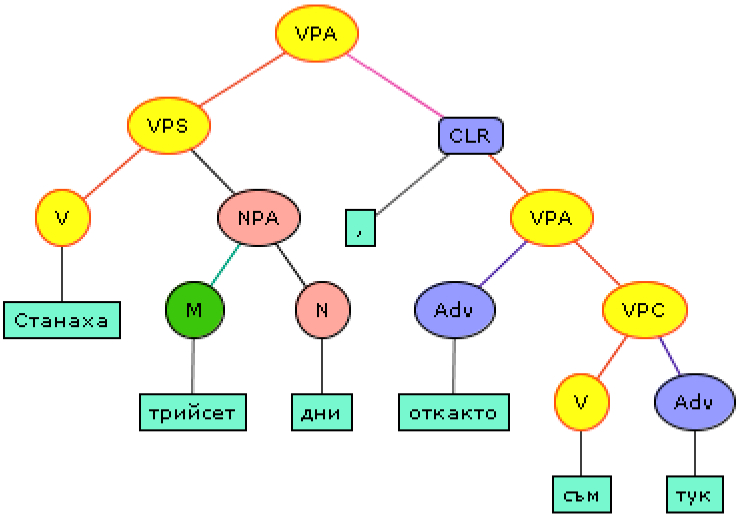
Subject



Complement



Adjunct



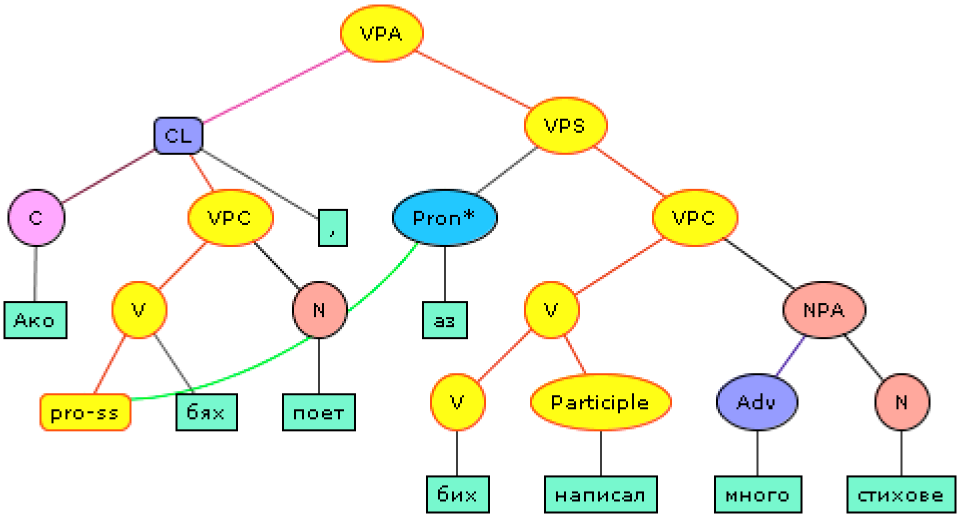
Concerning non-free relatives, they could be attached either to the whole *NPA*, when modifying its head, or to the governed noun itself, when modifying it only. When non-free relative is a subject or a complement then it is attached to some non-expressed element on the **ARG-ST** list of the verb lexical head. We do not annotate it here as a coreferential relation because such a relation can be easily inferred.

4.5.6 Other Types of Clauses — CL

CL element is used for any other clause which does not fit in the above subtypes. Here we give some examples of the most frequent clauses:

If-Clause

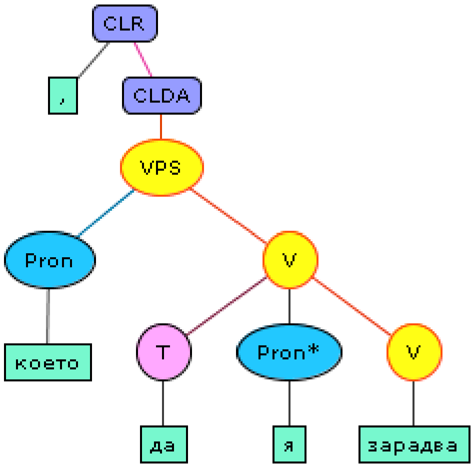
If-clauses exspress condition and are usually introduced by the subordinator 'ако' ('ako', if) and it is separated from the main clause by comma.



There are special types of *if*-clauses when the condition is not introduced by 'ако', but by other subordinatiors like 'когато' ('kogato', when-relative), but according to the guidelines it projects *CLR*. Also the interrogative particle 'ли' ('li') as well as the 'da'-form can have the meaning of 'ако' conjunction.

Combined Clauses

It often happens that one clause represents two types. It holds especially for 'da'-constructions with the presence of an interrogative or relative word. In such cases first the *CLDA* is marked and then — higher — the other one — *CLQ* or *CLR*. Here is one example:



4.5.6 Sentence Level Clause — S

The tag *S* corresponds to a saturated verbal phrase which is the highest node of the sentence. Recall that in cases where the full stop separates the head and the clause, the detached clause is marked *CL* or another appropriate type of clauses. Sometimes *S* is annotated to only fragments of sentences and then the top node is not a saturated verb phrase. However, sentences with direct and reported speech require special consideration. For more details see in BTB-TR02 report — [Simov and Osenova 2004a].

5 Noun Phrase

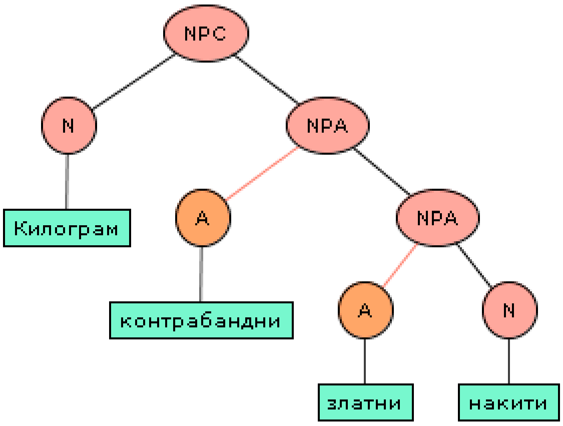
The nominal domain consists of the lexical element *N* and the phrasal elements: *NPC*(omplement), and *NPA*(djunct). In this section we present the phrasal elements in the domain. Following the hierarchy of the realization of the dependents we have this realization: *lexical noun* -> *nominal head-complement phrase* -> *nominal head-adjunct phrase*. The lexical head of the phrasal element can be also a pronoun, a family name, substantivized or nominalized element.

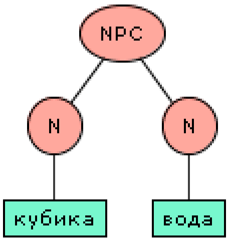
5.1 Head-Complement Noun Phrase (NPC)

NPC element corresponds to a noun *head-complement* phrase. Each phrase of this type must have a head daughter and a complement daughter. *NPC* can have only a lexical head and this includes the following elements: *N*, *Subst*, *Nomin*, *N-elip*, *ND-elip*, *CoordP*. We consider as *NPC* only nominal groups which generally represent the relation 'quantity-entity' (traditionally noted as an *NP NP* group). There are several subtypes of this general group. All other cases with relational nouns as heads are treated as nominal groups of a *head-adjunct* type. In an *NPC* group the first element is always the head. Thus the possible complements include nominals that can be measured. The head expresses the measure. The head can be a coordination phrase, substantivized, nominalized or elliptical. The subtypes of *NPC* are as follows:

Quantity-Substance

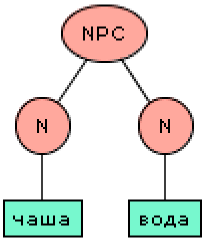
They consist of a noun for a measure and a noun for the material that is measured: 'литър мляко' (liter-m milk-n, litter of milk).

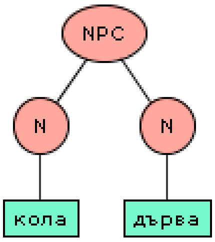




Container-Content

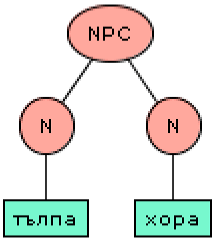
Phrases in which the head is a container of some sort and the complement is the content of the container: 'чаша вода' (glass-f water-f, (glass of water).

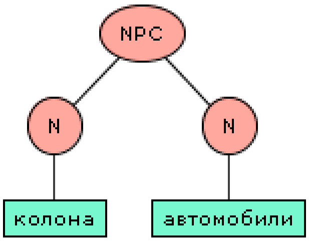




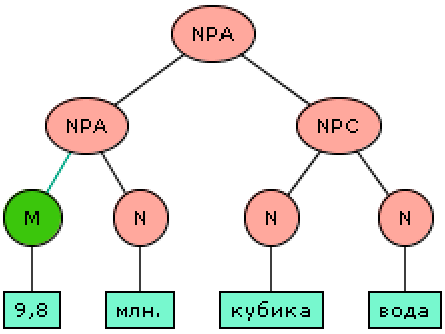
Type of Assembling-Entities

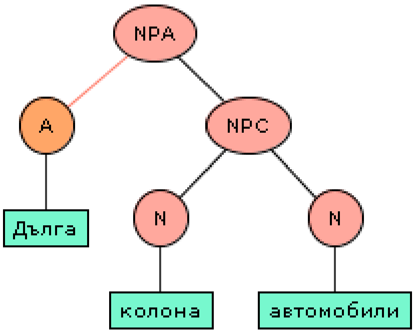
Here the head is a noun for a collection and the complement represents the members of the collection: група студенти (group-f,sg students-m,pl, a group of students), вид сирене (type-f,sg cheese-n,sg, a type of cheese).





Note that when there are premodifiers of the first noun, then they come last, because it takes its complement first. For example, we rely on the following analysis:





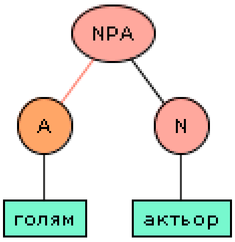
5.2 Head-Adjunct Noun Phrase (NPA)

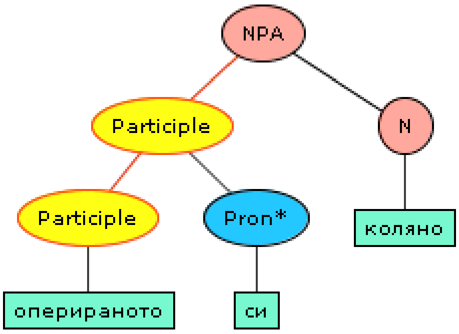
NPA element corresponds to a noun *head-adjunct* phrase. Each phrase of this type must have a head daughter and an adjunct daughter. *NPA* can have a lexical or phrasal head and this includes the following elements: *NPA*, *NPC*, *N*, *Pron*, *H*, *Subst*, *Nomin*, *N-elip*, *ND-elip*, *CoordP*. The possible adjuncts include nominals, numerals, adverbs, prepositional phrases, adverbials, clauses, participles, coordinated phrases. The adjuncts could be pre-modifying *AP*s (including some pronouns), post-modifying *PP*s, post-modifying clauses (*CLDA*, *CLQ*, *CLR*). The head is always easy to determine. In case of *NN* (or *NP NP*) groups the first element is always the head. The adjunct types are as follows:

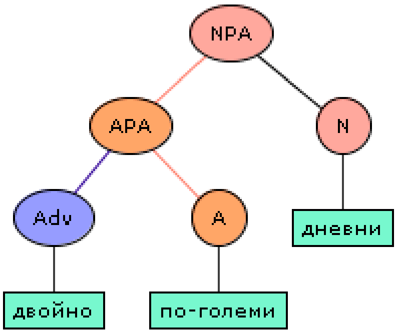
Adjectives, Participles, APs

There are several types of noun groups modified by adjectives, participles or adjective phrases.

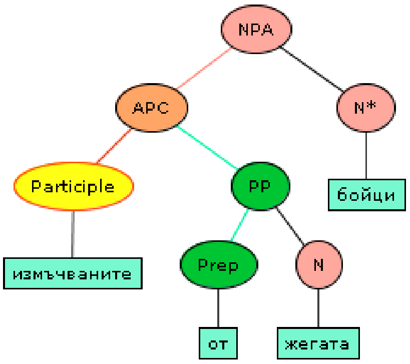
A noun (noun phrase) pre-modified by an adjective, a numeral, a pronoun, an adverb, a participle or an *AP*. This group is the basic one with respect to the pre-modifiers. Here are some examples:

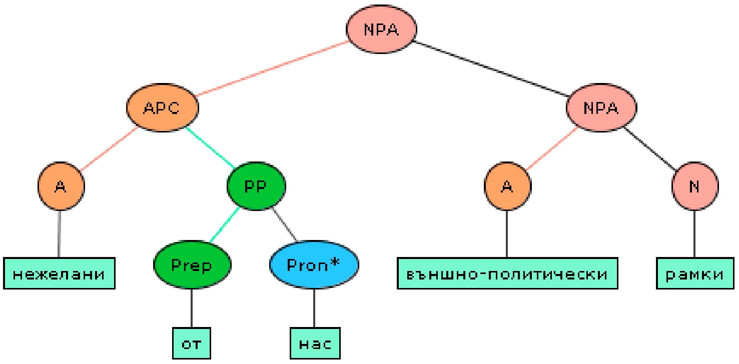




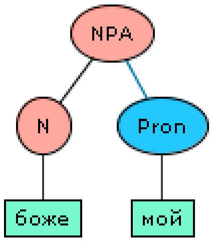


The first example shows the simplest case of a noun modified by an adjective. The second example shows a noun modified by a participle, which has a reflexive clitic attached to it. The last example demonstrates a noun modified by *APA*, which consists of an adjective modified by an adverb. The next two examples present a noun and *NPA* modified by *APCs*.

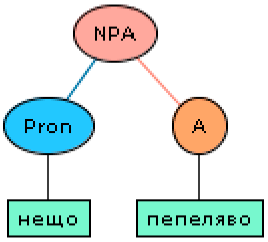


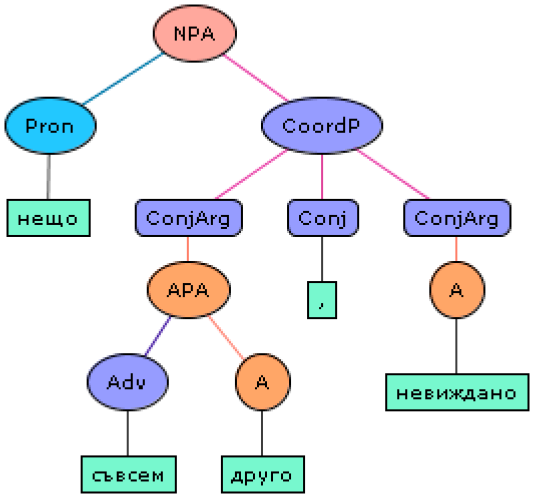


A vocative noun head can be pre-modified or post-modified by an adjective, a numeral, a pronoun:

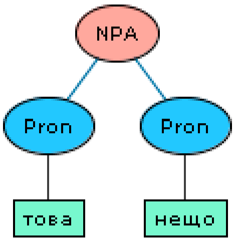


An indefinite, interrogative or negative pronoun can be post-modified by an adjective or an adjective phrase. Here are some examples. The second one demonstrates a complex modifier within a coordination phrase.



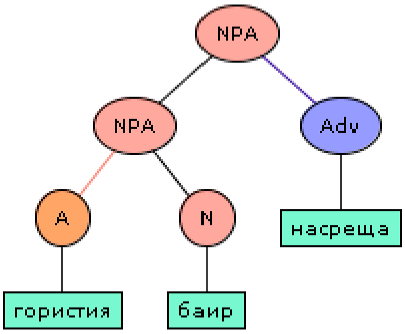


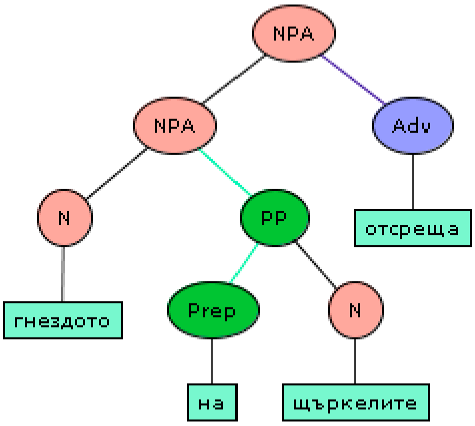
An indefinite or negative pronoun can be pre-modified by an adjective or a pronoun:



In these examples (and also in the previous ones) the head is determined on the basis of the morphosyntactic information of the nouns.

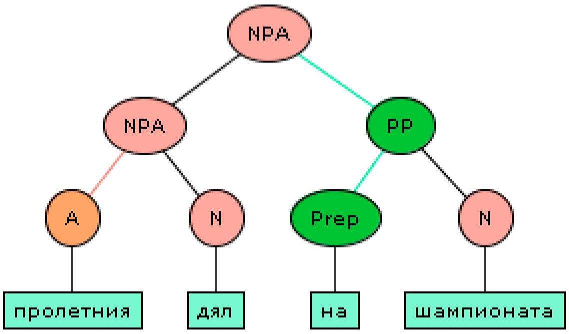
A noun or noun phrase can be post-modified by an adverb:

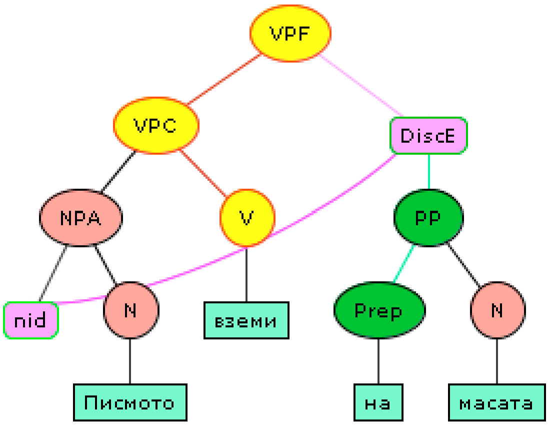




Prepositional Phrases

A noun or a noun phrase can be post-modified by a prepositional phrase - *PP*. In some cases the prepositional phrase can be extracted outside of the noun phrase and then its constituent position is represented locally as a *nid* element. It is accepted that the prepositional phrases are realized after the pre-modifying adjuncts. Here are some examples:





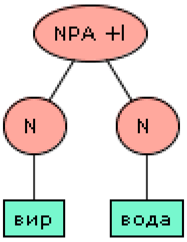
The first example shows the higher realization of the prepositional phrase. The second one demonstrates the extraction of a *PP* from an *NPA*. Recall that the *nid* element does not show the word order position of the element.

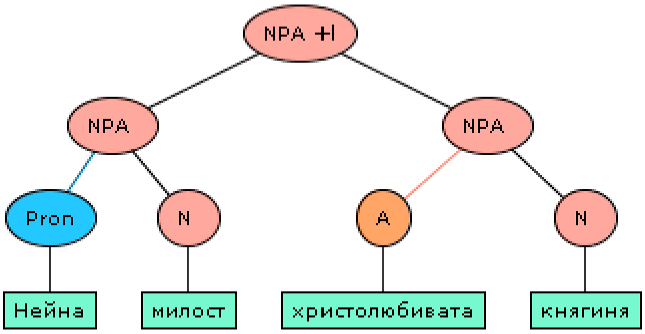
Noun and Noun Phrases Modifiers

1. Type of entity-qualification (apposition)

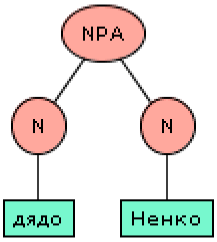
There are two types of such noun phrases. One, in which the adjunct is in preposition and another, in which it is in postposition. The first case is problematic with respect to the identification of the head. Up to now there was no problem to determine the head of the phrase on the basis of the phrase's type or the type of its constituents, their morfosyntactic features or the word order. In this case we need to mark up the head explicitly. This is done by introducing an attribute *head* for the *NPA* element. The attribute has three values. The first is *def* — default value. In this case the rules for determining the head are applied; the second is *f* — with the following meaning: the head is in first position; and the third is *l* — which means that the head is in the last position. Thus, the general rule for noun phrases having two nominals as daughters is that the first daughter is the head. When the modifying nominal is first and there are no other applicable rules, we annotate the NPA element with the attribute *head* with value *l*.

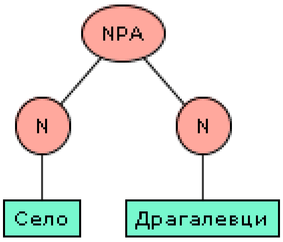
There are two main kinds of phrases with a modifying nominal in first position: (1) phrases like 'злато жена' (zlato zena, gold-n lady-f, a lady like gold), and (2) 'дядо Петър' (djado Petar, grandfather-m Peter-m, grandfather Peter), 'професор Тодоров' (profesor Todorov, professor Todorov). The first type is relatively rare, although still productive ('огън кафе', coffee like a fire, taken from an advertisment). The second type follows the pattern: a noun for relative relation, profesion, or title followed by a personal name. Because the second pattern is easy to recognize we do not annotate it with the attribute *head*. In all other cases we annotate the head. Here are some examples. The head attrbute is represented as a plus sign after the element label and the value is after the plus sign.

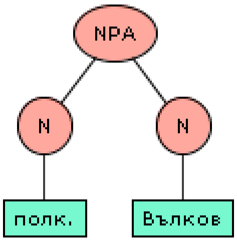




These two examples show the problem with the identification of the head. The attribute *head* says that the second constituent is the head in both cases.

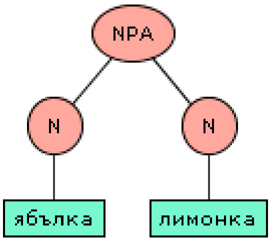


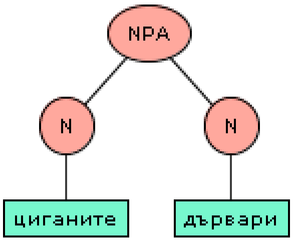




The above examples demonstrate a proper noun with a nominal modifier. Here we apply the rule, which says that the proper nominal is the head. The proper nominal could be a noun phrase as well.

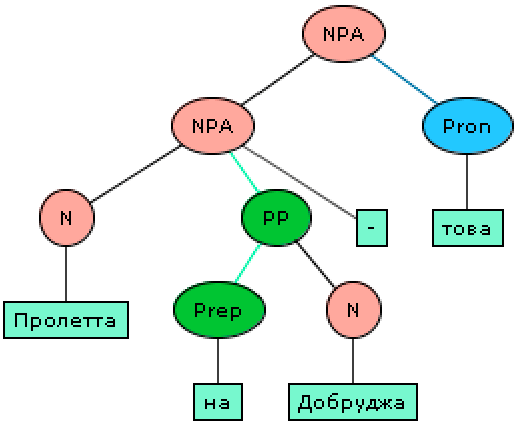
Here are some examples of phrases in which the modifier is in the second position. The head always comes as a first constituent.





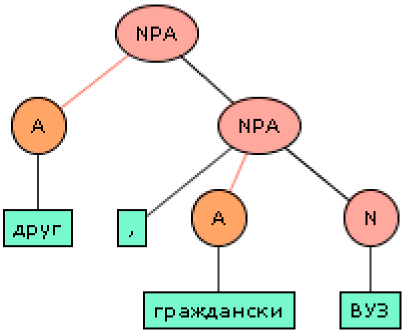
2. Type of a head-pronoun which is modified by a noun phrase

Here the head is determined by the POS: a pronoun. In the example below the demonstrative pronoun 'това' (‘this’) is modified by an NPA phrase. In this way we express such types of nominalization:



Pre-positive Modifiers Separated by Comma

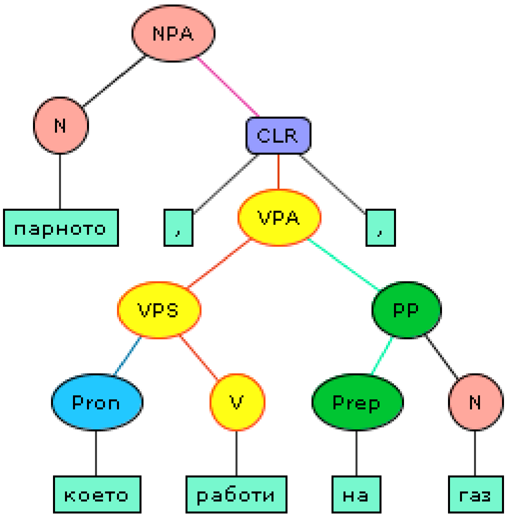
This is a type of pre-positive modifiers separated by comma, which is different from the coordination of two modifiers.

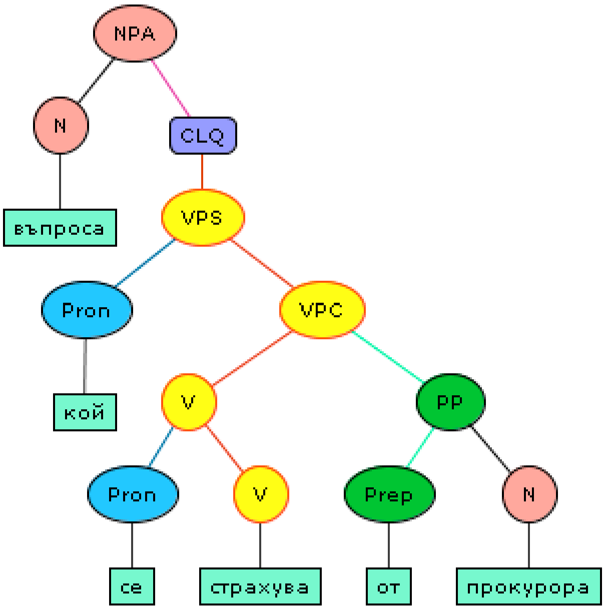


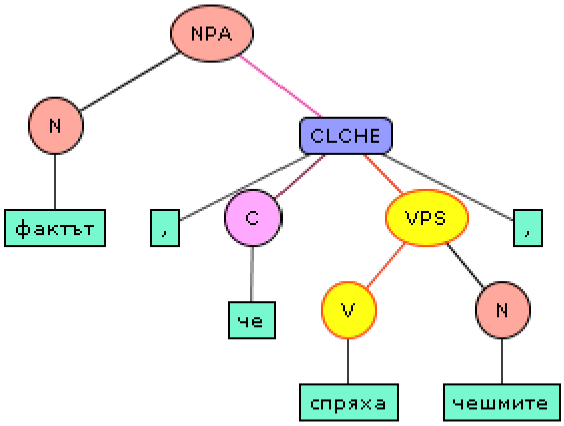
In the above example the two modifiers are taken as adjuncts, i.e. one by one. This relation between two modifiers, separated by a comma, is typical for coordination. However, here it expresses adjunction. The difference is distinguished only on semantic grounds.

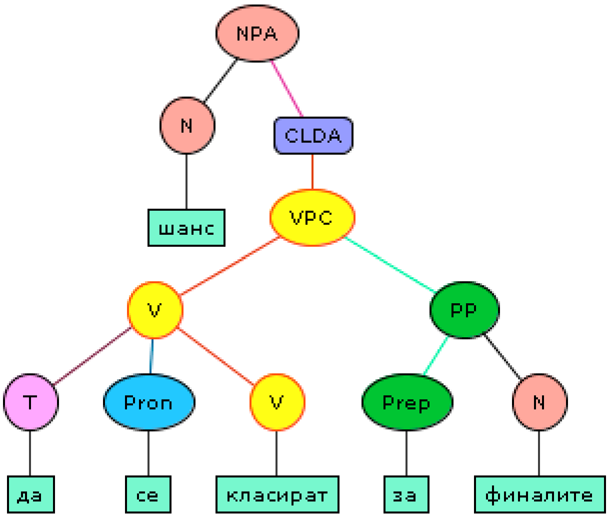
Clausal Postmodifiers

The clausal post-modifiers are usually *CLR*, *CLQ*, *CLDA*, and *CLCHE*. They can modify either the head of a noun phrase, or its modifiers. Below there are examples for each kind of clauses. The first example is a noun modified by a relative clause. The relative pronoun is the head of the first constituent in the clause and it is always co-referent with the head noun of the noun phrase. This co-referential relation is not explicated, because it can be easily inferred. The second example shows a noun modified by an interrogative clause. The third and the fourth examples demonstrate nouns modified by 'da'-clause and 'che'-clause, respectively. In all three cases (i.e. except for the relative clause) the predominant interpretation is that the head noun of the noun phrase expresses a nominalization of the clause. Such a nominalization is represented by words for cognitive facts like 'въпрос' (question), 'мисъл' (thought), мнение (opinion), факт (fact), желание (wish) etc. This nominalization is not explicated. Only when these types of clauses have constituents which are co-referent with the elements of the noun phrase they modified, then we express these relations overtly. Very often these clauses are used directly as arguments of the verbs and then they modify the unexpressed nominal referents in the **ARG-ST** list of the verb. For such examples see the section on clauses above.



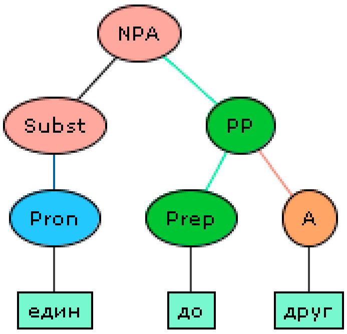


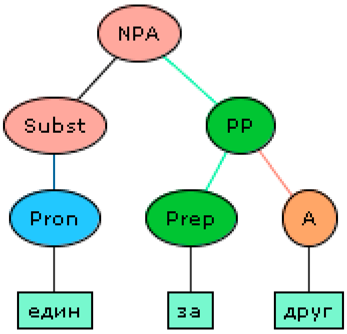


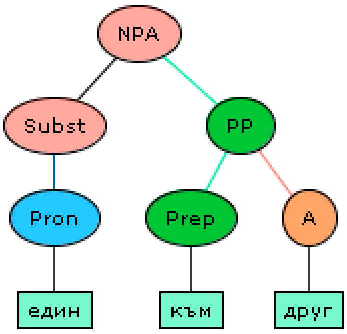


Reciprocal NPA

We treat reciprocals as *NPAs* with a substantivized head, modified by a PP. The reciprocals are always co-referred with the subject or the object of the sentence. Here are some examples:

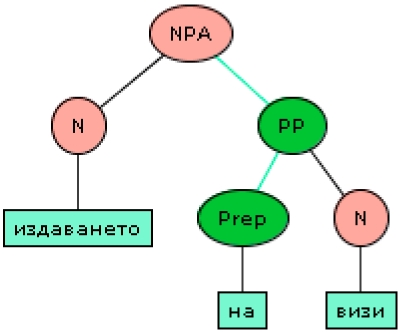


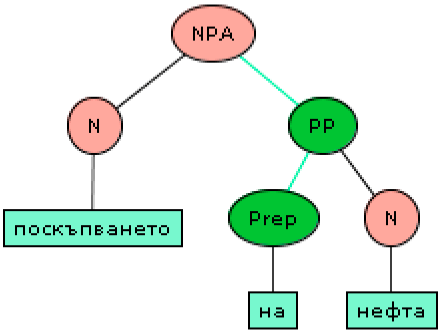




Deverbal Nouns

Deverbal nouns inherit the **ARG-ST** list from verb and most of the *PP* modifiers correspond to the arguments of the verb, but their presence in the noun phrase is much less required than in the verb phrase. This is why we treat them as modifiers and hence, the noun phrases with a head deverbal noun are *NPA*. Here are some examples:





6 Adjective Phrase

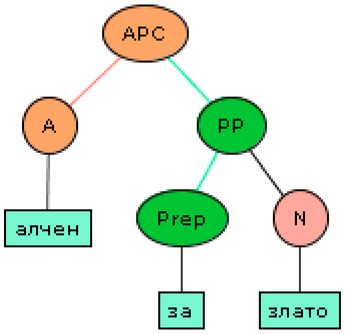
There two types of adjective phrasal elements: *APC*(omplement) and *APA*(djunct). Following the hierarchy of the realization of the dependents we have the following realization: *lexical adjective* -> *adjectival head-complement phrase* -> *adjectival head-adjunct phrase*. The lexical head of the phrasal element can be also an adjectival pronoun, a family name adjective, or a participle element.

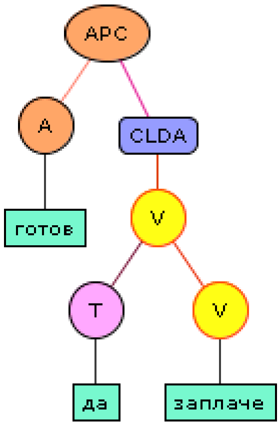
6.1 Head-Complement Adjective Phrase (APC)

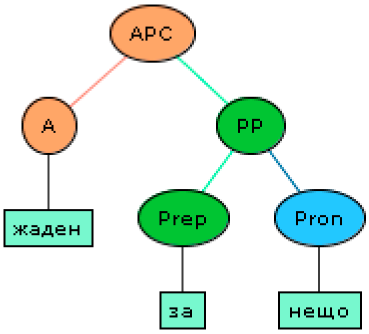
APC element corresponds to an adjective *head-complement* phrase. Each phrase of this type must have a head daughter and complement daughters. *APC* can have only lexical head and this includes the following elements: *A*, *Participle*, *CoordP*. This distinction is important, because both types (adjectives and participles) of heads behave differently. The *participle* as a head inherits all the arguments of the corresponding verb and therefore, the determination of its possible complements depends straightforwardly on the corresponding verb. The *adjective* as a head relies on a list of relational adjectives, which sometimes cannot be determined clearly. The identification of the head becomes a problem when both - the head and the complement are participles. In this case we apply the same rules as in the case of *VPC*. The subtypes of *APC* are as follows:

Adjective with Complements

There are adjectives that require complements: 'алчен за' (‘alchen za’, greedy for), 'готов на/да' (‘gotov na/da’, ready for), 'жаден за' (‘zhaden za’, eager for). Usually the complement is a prepositional phrase or a clause. Here are some examples:

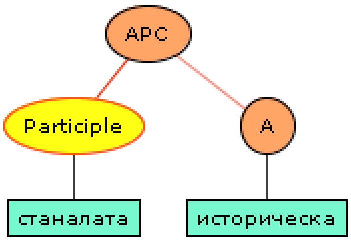


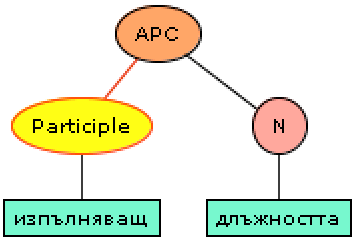


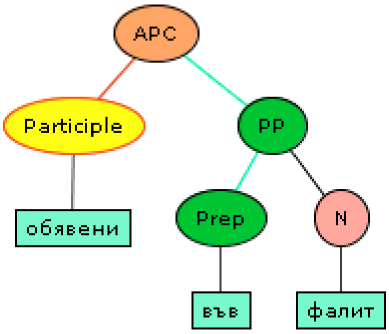


Participle with Complements

As it was mentioned above, the participles inherit the arguments of the verb and when they are used attributively or predicatively, they can project an *APC* with appropriate complements. In some cases a complement can be extracted (usually in the predicative usage) and then its constituent position is represented by the *nid* element. Note that the subject of the verb is treated as an adjunct when subcategorized by a participle, i.e. as an *APA* phrase (see below). Here are some examples of *APCs:*

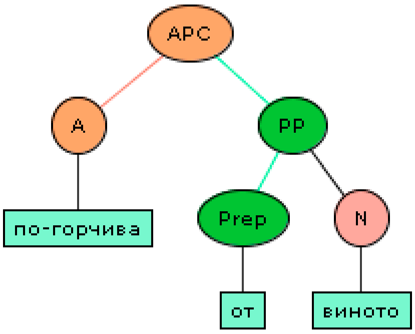


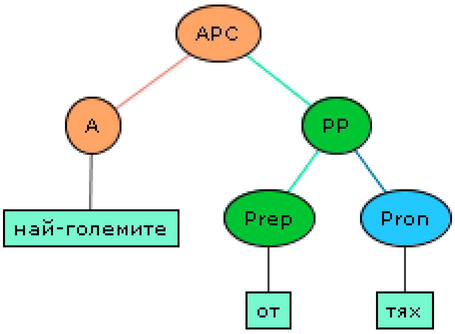




Comparative and Superlative Forms

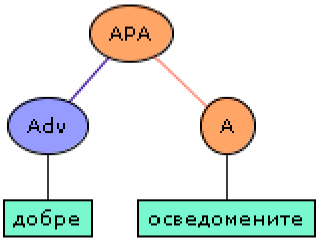
The comparative and superlative forms of the Bulgarian adjectives and participles require a specification of the entities, which they evaluate with respect to some characteristics. We accept that these evaluated expressions are complements. They are usually prepositional phrases with prepositions 'от ('ot', from), 'сред' ('sred', among) etc. Here are some examples:

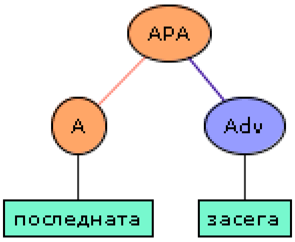


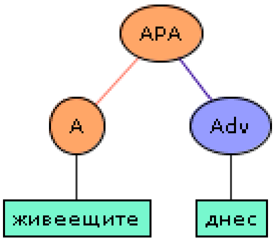


6.2 Head-Adjunct Adjective Phrase (APA)

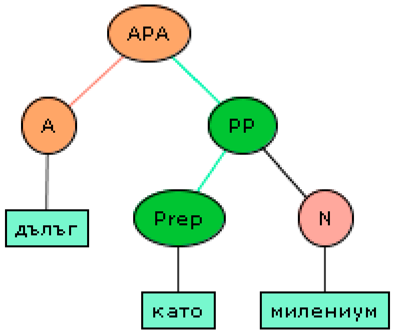
APA element corresponds to an adjective *head-adjunct* phrase. Each phrase of this type must have a head daughter and an adjunct daughter. *APA* can have a lexical or phrasal head and this includes the following elements: *APA*, *APC*, *A*, *Participle*, *H*, *CoordP*. The possible adjuncts include adverbs, nominals, prepositional phrases, clauses, coordinated phrases. The adjuncts could be pre-modifying or post-modifying adverbs (or adverbial phrases), post-modifying clauses. If the head is a participle, then all possible adjuncts for the verb can appear as adjuncts in the adjective phrase. The head is always easy to determine. The most frequent case of adjuncts in the *APA* element are adverbials. The adjunct can be in pre- or postposition. Here we give a few examples:

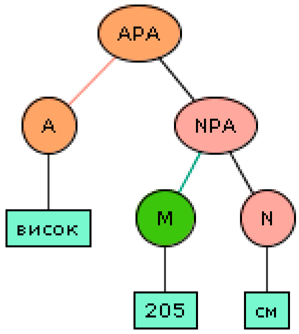


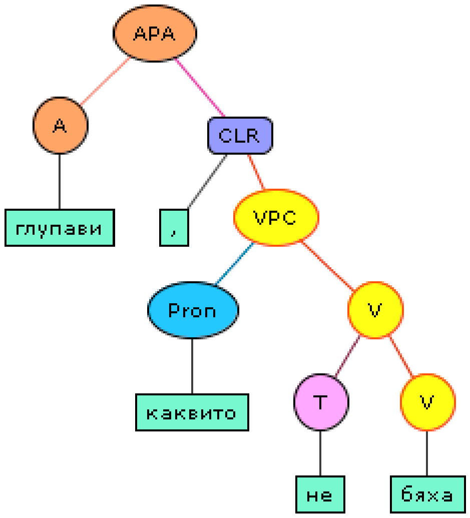




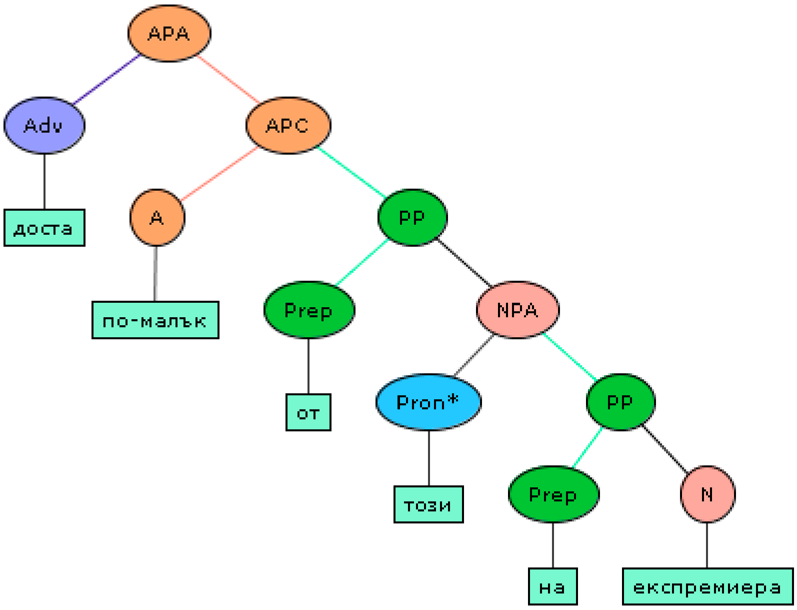
Other types of adjuncts can include clauses, nominals, prepositional phrases. Here are some examples:







There are also cases in which the head of the *APA* is *APC*. In the example below the comparative adjective takes first its *PP* complement and then the prepositive *Adv* modifier:



7 AdverbPhrase

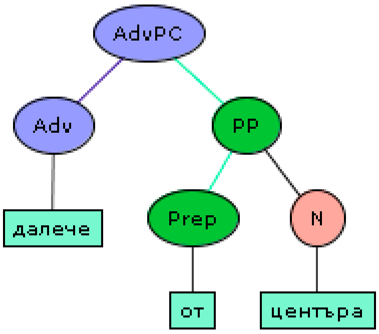
There two types of adverb phrasal elements: *AdvPC*(omplement), and *AdvPA*(djunct). Following the hierarchy of the realization of the dependents we have the following realization: *lexical adverb -> adverbial head-complement phrase -> adverbial head-adjunct phrase*. The lexical head of the phrasal element can be also a pronoun adverb, a gerund element.

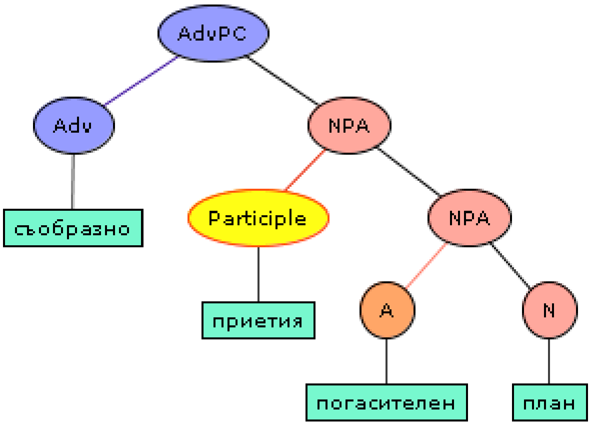
7.1 Head-Complement Adverb Phrase (AdvPC)

AdvPC element corresponds to an adverb *head-complement* phrase. Each phrase of this type must have a head daughter and complement daughters. *AdvPC* can have only lexical head and this includes the following elements: *Adv*, *Gerund*, *CoordP*. Because we consider that the gerund inherits all the arguments of the corresponding verb, the possible complements can be a wide range of elements. The identification of the head is not problematic. The subtypes of *AdvPC* are as follows:

Adverb with Complements

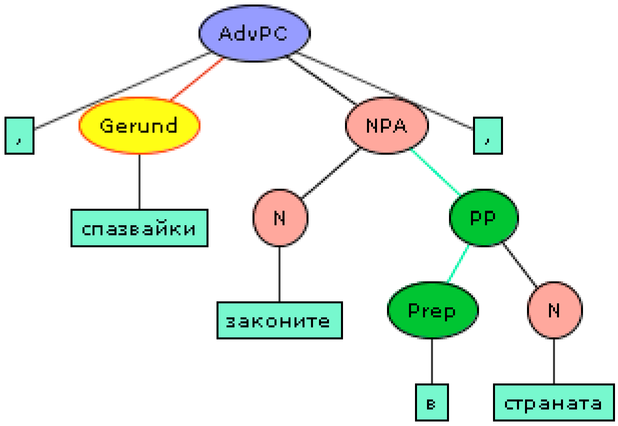
Some adverbs require arguments. Here are two examples:

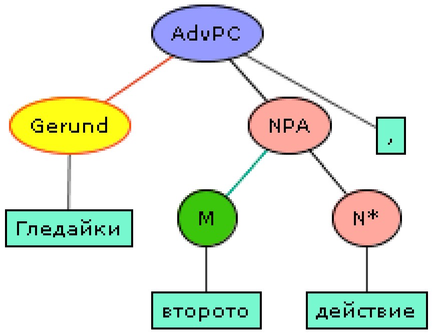




Gerund with Complements

As it was mentioned above, the gerunds inherit the arguments of the verb and they can form *AdvPC* with appropriate complements. Here are some examples:

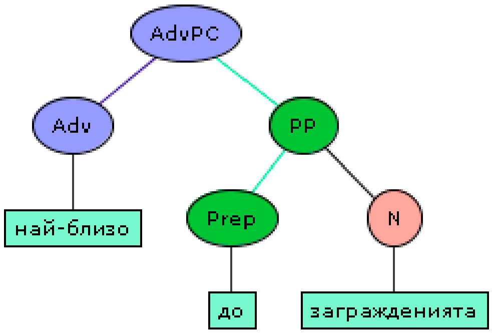


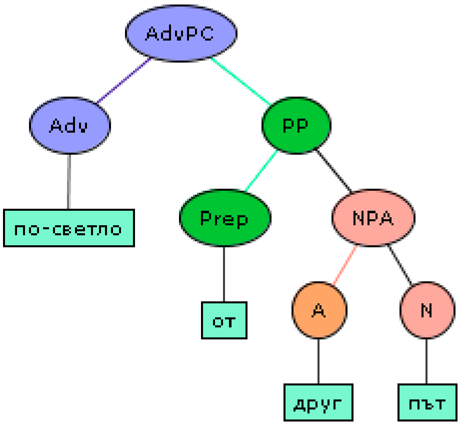


The phrases of this kind are separated from the rest of the sentence by comma(s). The inherited from the verb subject is always co-referential with the subject of the main clause. Because this co-referential relation is easily inferable we do not represent it explicitly.

Comparative and Superlative Forms

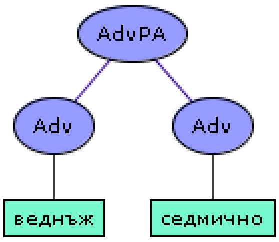
Similarly to the adjectives, the comparative and superlative forms of the adverb require evaluated expressions as a complement. Here are some examples:

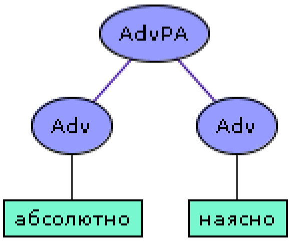


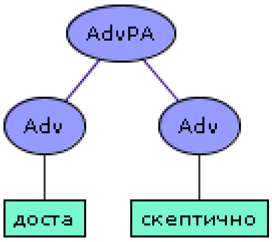


7.2 Head-Adjunct Adverb Phrase (AdvPA)

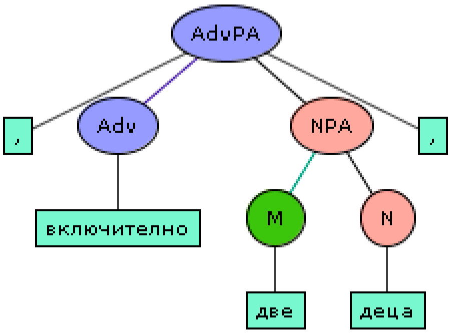
AdvPA element corresponds to an adverb *head-adjunct* phrase. Each phrase of this type must have a head daughter and an adjunct daughter. *AdvPA* can have a lexical or phrasal head and this includes the following elements: *AdvPA*, *APC*, *Adv*, *Gerund*, *CoordP*. The possible adjuncts include adverbs, nominals, prepositional phrases, clauses, coordinated phrases. If the head has a gerund as a lexical head, then all possible adjuncts for the verb can appear as adjuncts in the adverb phrase. The head is always easy to determine. The most frequent case of adjunct in *AdvPA* element is adverbials. In this case the adjunct is always in preposition. Here we give a few examples:

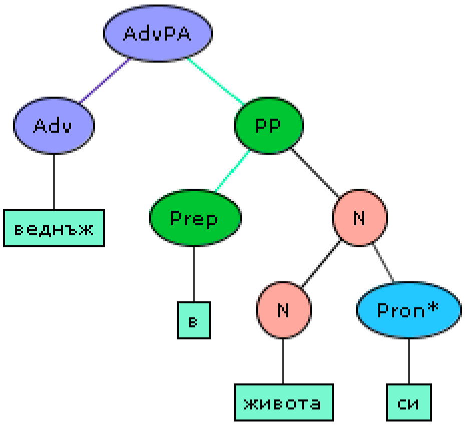


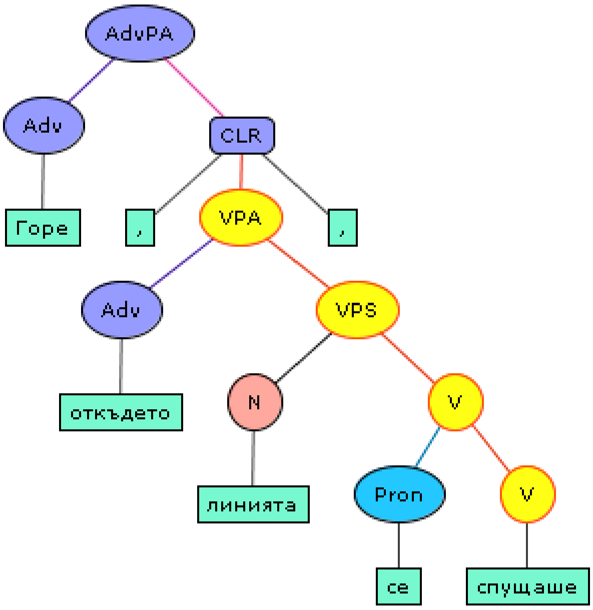




Other types of adjuncts can include clauses, nominals, prepositional phrases. Here are some examples:





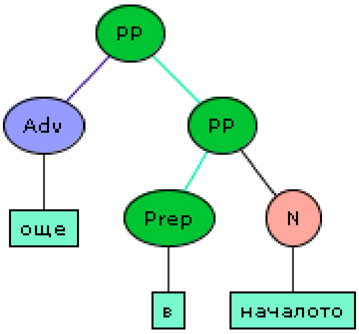


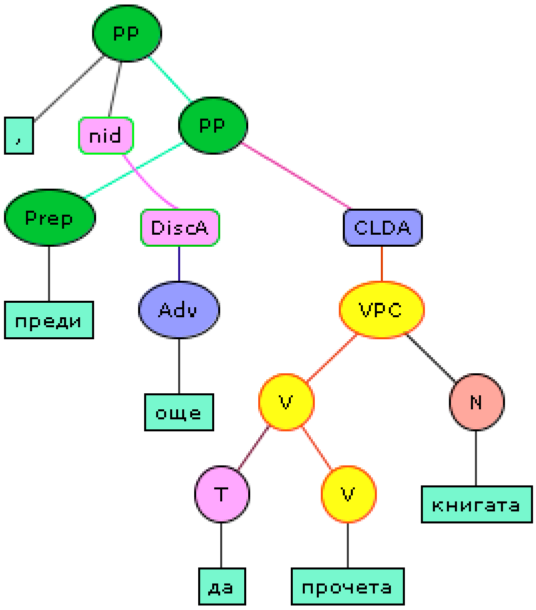
8 Prepositional Phrase

PP corresponds to a prepositional *head-complement* and *head-adjunct* phrase. To be more precise, there are two types of *PP*s: of complement type - *PPC*(omplement) and of adjunct type — *PPA*(djunct), but as the latter are rare, we do not make this distinction explicit. Thus, all prepositional phrases are represented as a *PP* element and the prepositional head adjunct phrase are recognized on the basis of their internal structure, see below.

PPs with Adverbial Adjuncts

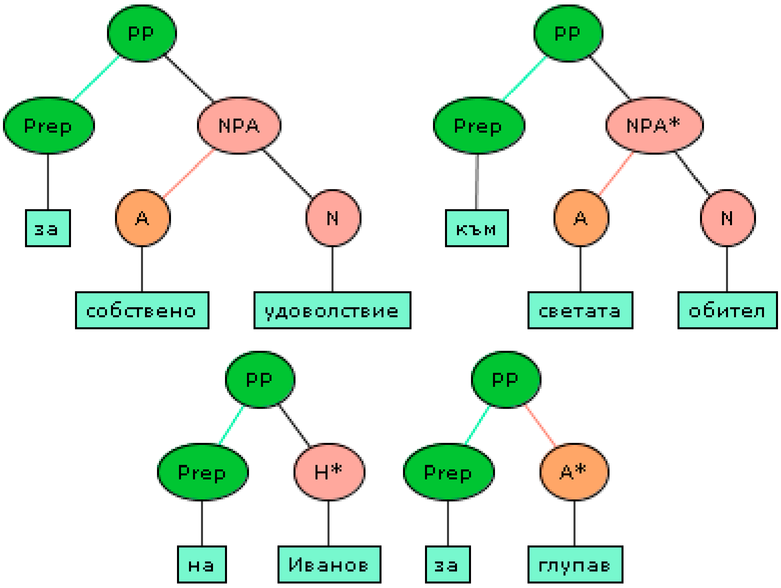
These are prepositional phrases which comprise a *PP* element and an emphatic adverb (typical adverbs are: още (oshte, yet), само (samo, only), etc). The head is always the *PP* element. The adjunct can be realized inside the *PP* head and then it is marked-up with a *DiscA* element. Here are two examples:





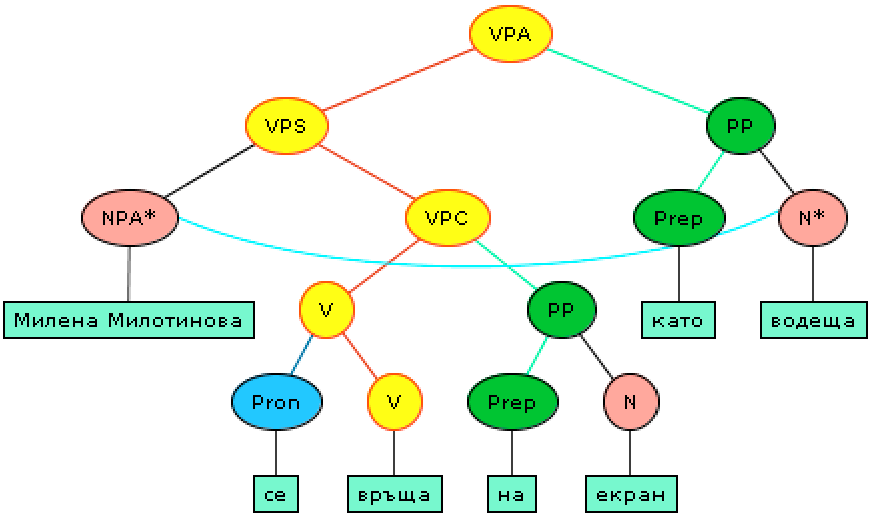
PPs with Nominal Complements

This is the most frequent type of prepositional complement. The head is a preposition and it might be elliptical — *Prep* and *Prep-Elip* elements. The complement can be: *N*, *H*, *A*, *APA*, *Pron*, *Nomin*, *Subst*, *NPA*, *NPC, CoordP* element. The complement is always after the head. Recall that the asterisks symbol means that this constituent is involved in a co-reference relation within the sentence. Here are some examples:

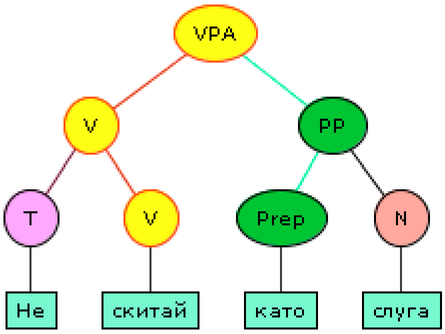


There are special *PP*s with a nominal complement headed by the prepositions 'като' (kato, as) and 'за' (za, for). When these prepositional phrases express secondary predication, their *NP* complements co-refer with the subject or object. When they express adjunct of comparison, then no co-reference is established.

Here is an example for secondary predication:

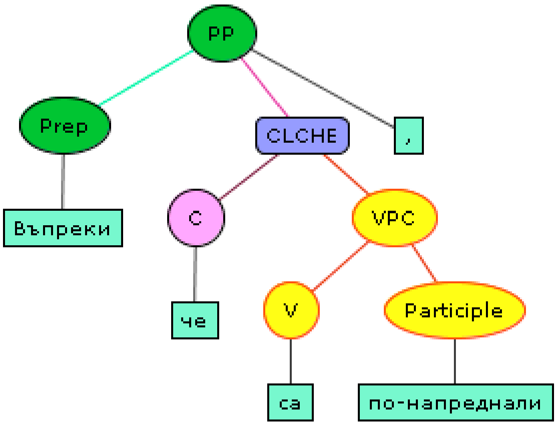


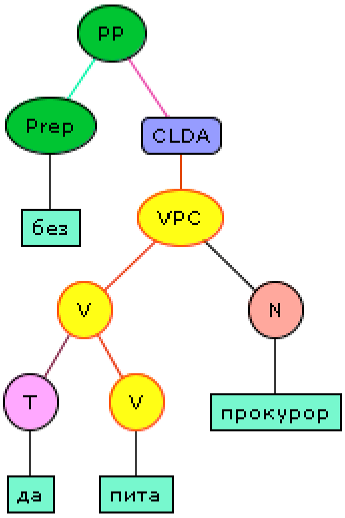
Here is an example for adjunct of comparison:



PPs with Clausal Complements

The following prepositions can take clausal complements (either *CLDA*, or *CLCHE*, or both): 'преди' ('predi', before), 'освен' ('osven', except), 'вместо' ('vmesto', instead of), 'без' ('bez', without), 'въпреки' ('vapreki', although). A special case is the word: 'макар'. Note that the traditional grammar analyzes all of them as complex conjunctions when they head a clause. Here are two examples:



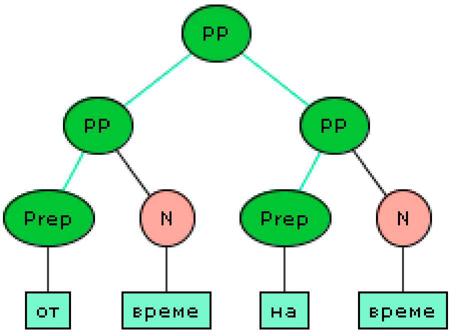


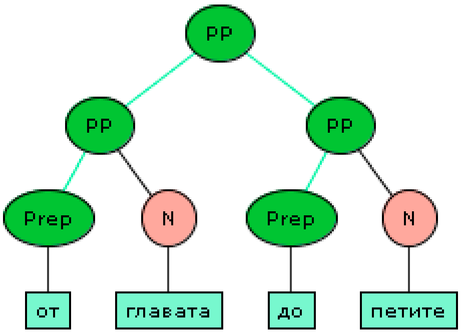
In the first example the preposition takes a *CLCHE* clause, while in the second one it takes a *CLDA* clause.

PPs with PP complements

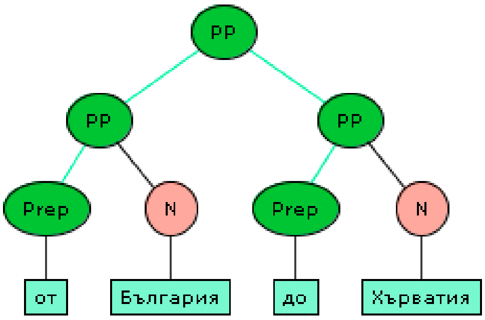
Here we include expressions for range. Usually the range is temporal or spatial interval. Because the complements in these cases are similar to the temporal or spatial adverbs, they can be also expressed by adverbials. We distinguish between the following types:

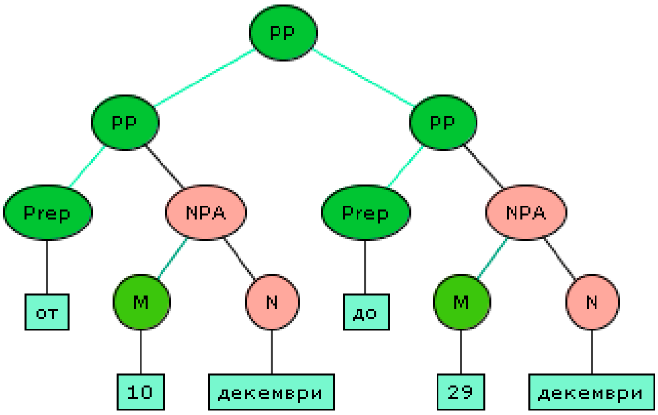
1. idiomatic expressions like: [PP [PP от време] [PP на време]] ('ot vreme na vreme', from time to time), [PP [PP от главата] [PP до петите]] ('ot glavata do petite', from head to toe), etc. Here are some examples:

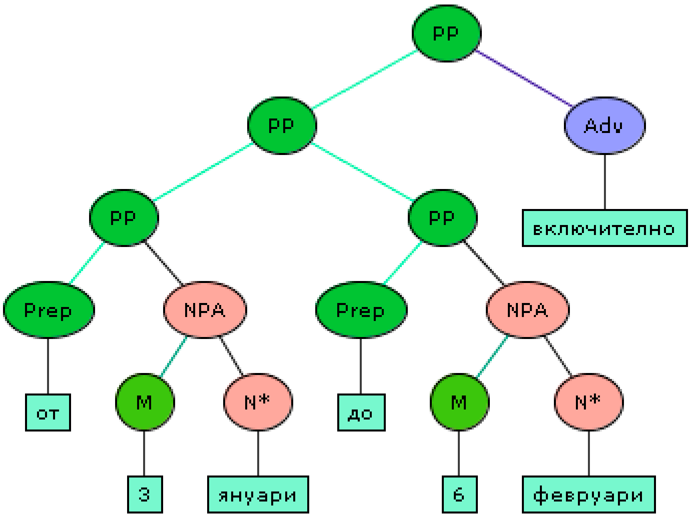




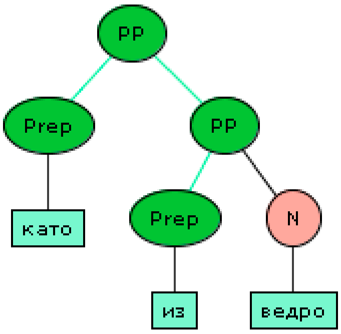
2. temporal and space expressions: [PP [от Варна] [до Бургас]] ('ot Varna do Burgas', from Varna to Burgas), [PP [от 6 юни] [до 20 август]] ('ot 6 yuni do 20 avgust', from 6 June to 20 August), etc. If the temporal expressions are modified by the adverb включително ('vklychitelno', including), then the adverb is attached to the highest *PP*. Note that these expressions can be realized one by one, if separated in the sentence (see in Preference rules section). Here are some examples:

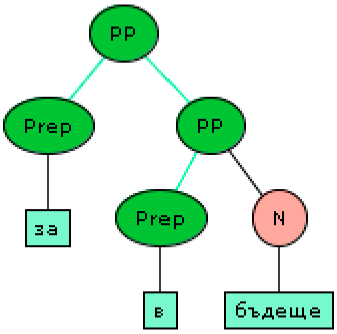


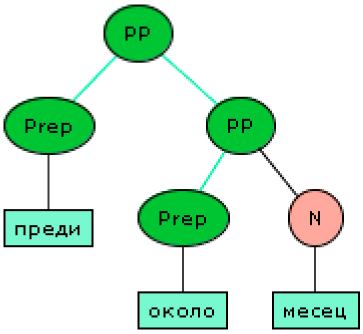




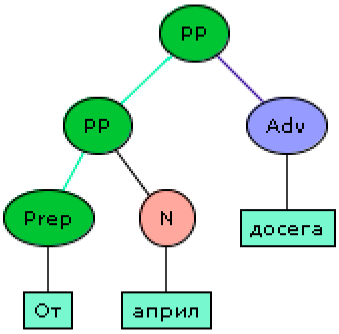
3. *PP*s with one *PP* complement: [PP като [PP за теб]] ('kato za teb', as for you).

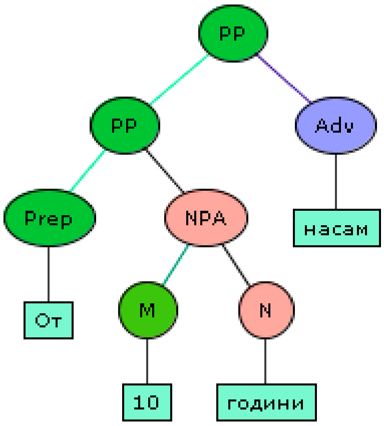






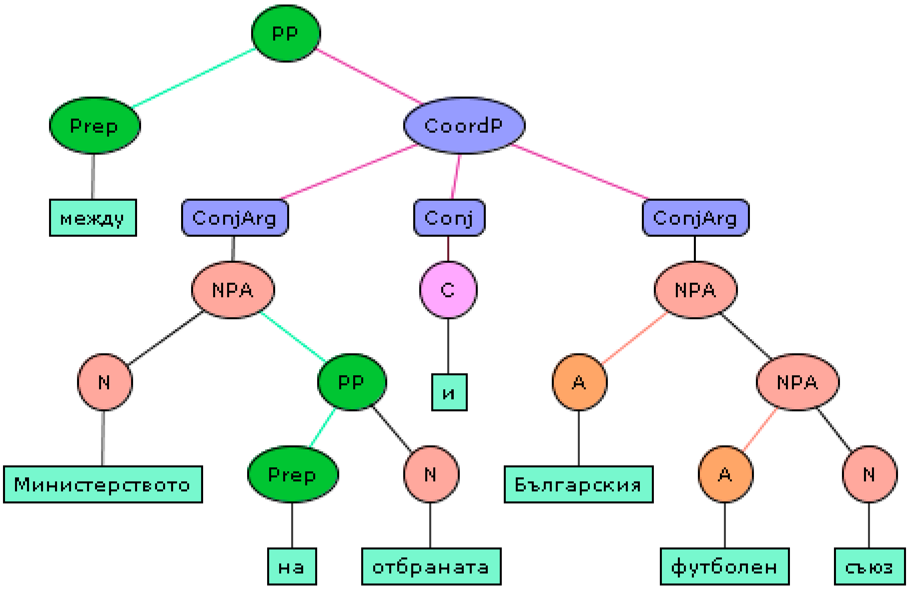
4. *PP*s with adverbial complements: The first type expresses the same as 'temporal and space expressions' above. Example: [PP [от април] [досега]] ('ot april dosega', from April till now). Here are more examples:





PPs with Collective Complement

These are prepositonal phrases with prepositions like между, измежду (between, among). They require a complement denoting more than one object. Very often the complement is a coordination phrase like in the example: Банката се намираше [PP между [CoordP пощата и театъра]] ('Bankata se namirashe [PP mezhdu [CoordP poshtata i teatyra]]', the bank was located between the post office and the theatre). Here is an example:



9 Coordination Phrase

CoordP corresponds to a *non-headed* phrase. The coordination phrase in our analysis has a flat structure. As it was mentioned earlier, it has only functional children which mark-up the role of each constituent: an argument of the coordination or a conjunction. The *ConjArg* element represents an argument (conjunct) of a coordination phrase; the *Conj* element marks-up the conjunction of the coordinated phrase together with the obligatory comma for the conjunction, if required. The underlying idea behind our treatment is that conjuncts within coordination have to agree in their grammatical function in spite of their syntactic category. We assume that coordination has to be treated as a *non-headed phrase* with the following requirements:

- The conjuncts have to agree in their valence potential: ***VALENCE** lists, **MOD**, and **SLASH** Feature*
- They can be underspecified with respect to the category: extension of the head value hierarchy

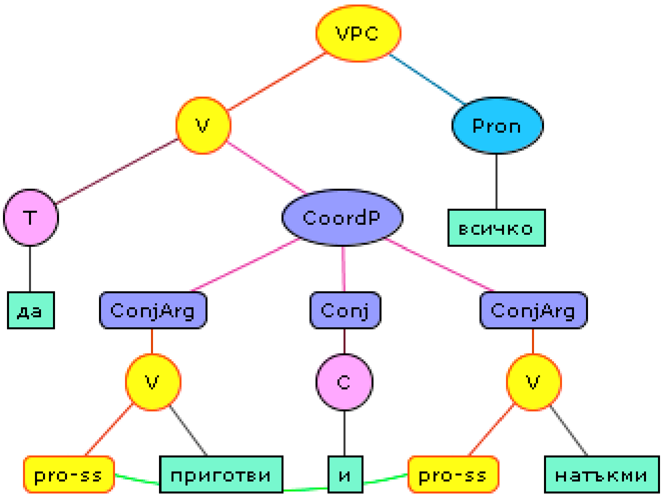
In this respect we propose not to classify obligatorily each coordination as an NP coordination, a VP coordination, etc., but to classify the coordination as *saturated coordination, adjunct coordination* and *unsaturated coordination.* Saturated coordination has empty valence lists, and the **MOD** feature has value *none*. Adjunct coordination has empty valence lists, and the **MOD** feature has value different from *none*. Unsaturated coordination has at least one non-empty valence list. In most of the cases the non-conjunction daughters of the coordination may also share the values of their other head features, but in some cases they disagree on them. In order to account for such cases, we changed the sort hierarchy by introducing the distinct sort *coordination*, which is at the same level in the sort hierarchy as *noun*, *verb*, *adj*, and *prep* sorts. In this way we underspecify the head of the coordination. Here is the modified sort hierarchy:

head
substantive (subst)
PRD : *boolean*
MOD : *mod-synsem* *none* for not adjuncts
coordination
noun
CASE : *case*
verb
VFORM : *vform*
AUX : *boolean*
INV : *boolean*
adj
adv
prep
PFORM : *pform*

This change in the sort hierarchy has also impact on the definition of selectional constraints in the lexicon.
We distinguish between the following types of coordination:

Lexical coordination

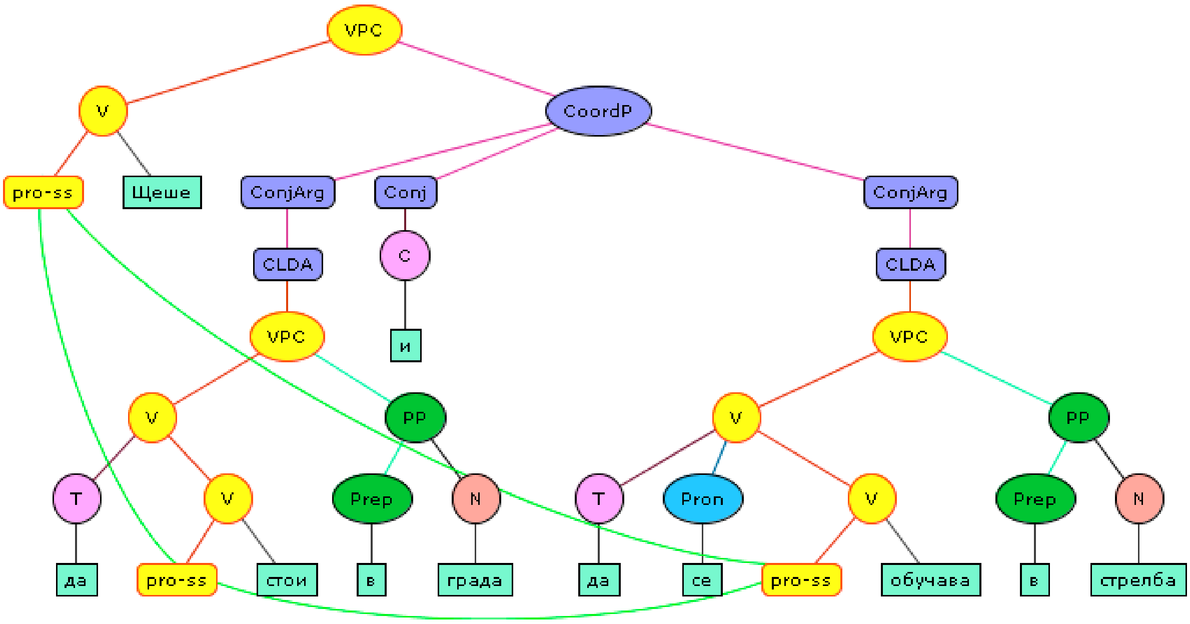
Typically each lexical category in Bulgarian (excepts for particles, interjections, clitics) can be coordinated on lexical level. This means that the coordination is a lexical sign, which has the same values for its **VAL** and **MOD** features as the coordinated lexical signs. Here we present an example of verbal lexical coordination:



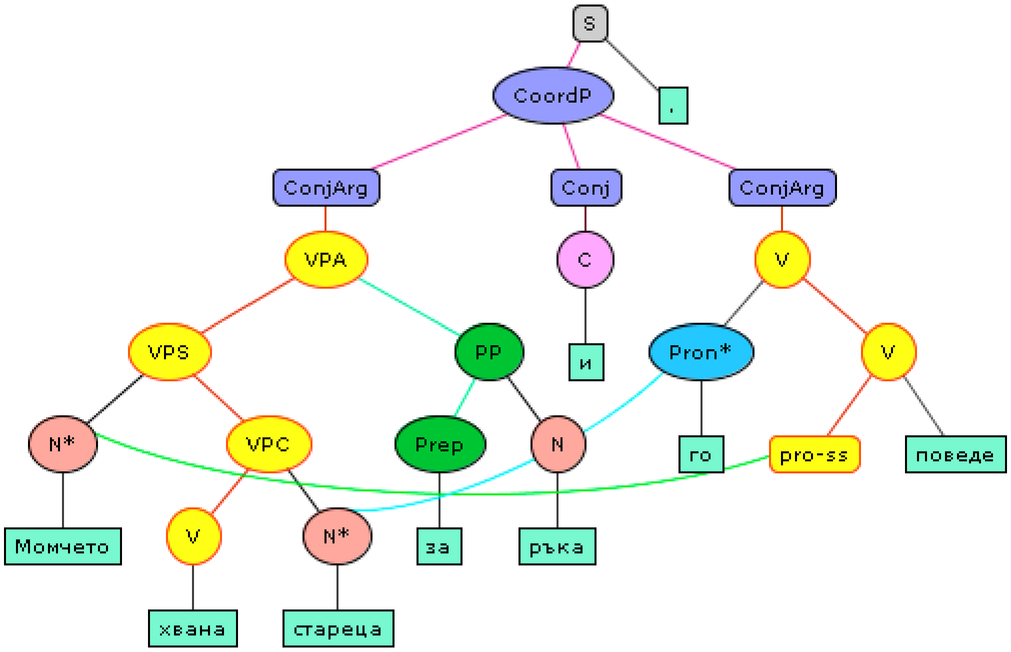
In the above sentence we have the special 'da'-construction in Bulgarian, where (as it was presented in the section on lexical verb) the verbal particle da takes a lexical verb and forms again a lexical sign, which inherits the **ARG-ST** list of the verb. In this case the two verbs agree in their **ARG-ST** lists (**MOD** feature in this case has empty values) and the coordination is possible at this level. Thus the coordination is 'приготви и натъкми' ('prigotvi i natykmi', she prepared and adjusted).

Sentential and Clausal Coordination

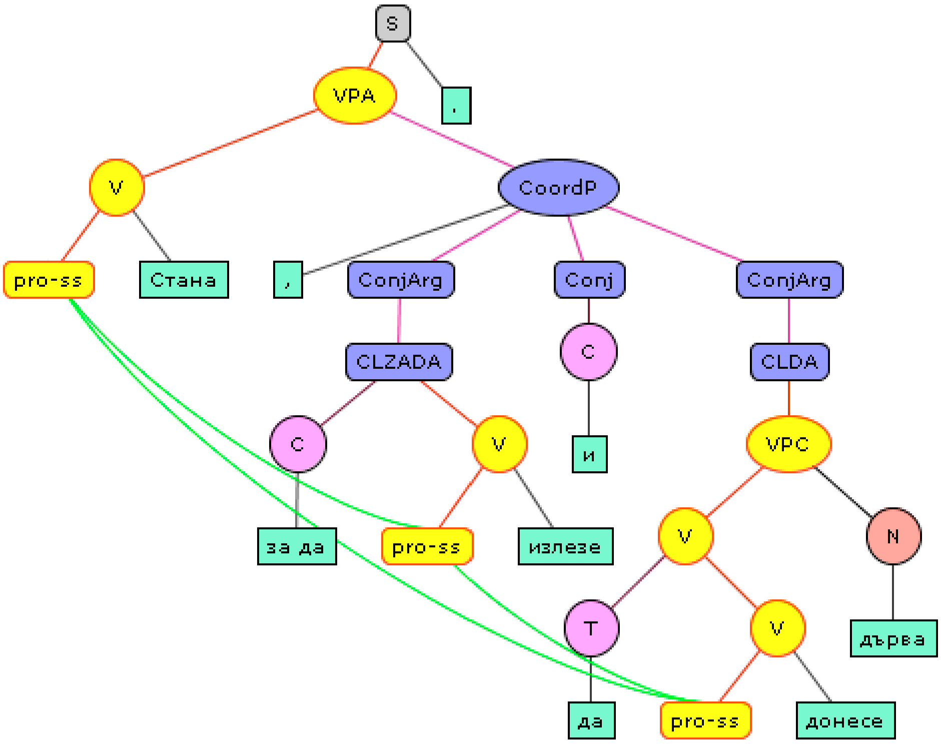
Modal and some auxiliary verbs in Bulgarians select for saturated *VP*s (clauses). Here the coordination is not a problem because the *VP*s are saturated and their valency lists are empty. Thus they trivially agree in their valency lists. Also they are selected by some of the head features like clausality and the type of the clauses. In the following example we again have da construction of two verbs, but they do not agree on their **ARG-ST** lists and consequently - on their valency lists. Thus the construction, possible in the example above, is not possible here. Because the coordination is selected by the verb 'щеше' (steshe, would), it means that the two coordinating phrases have unexpressed subjects and thus they are saturated *VP*s. The unexpressed subjects are co-referred with the unexpressed subject of the verb 'steshe' — the green curves:



In the next example, we have a sentence, in which the two verbs have identical valence lists, but the first one also has an adjunct 'за ръка' ('za ryka', by the hand), which has to be realized after the realization of the subject. For this reason it is a coordination of two sentences (clauses, saturated *VP*s). The second sentence has an unexpressed subject that is co-referred with the subject of the first one. There is a co-referential relation between the complement in the first clause and the complement clitic in the second.



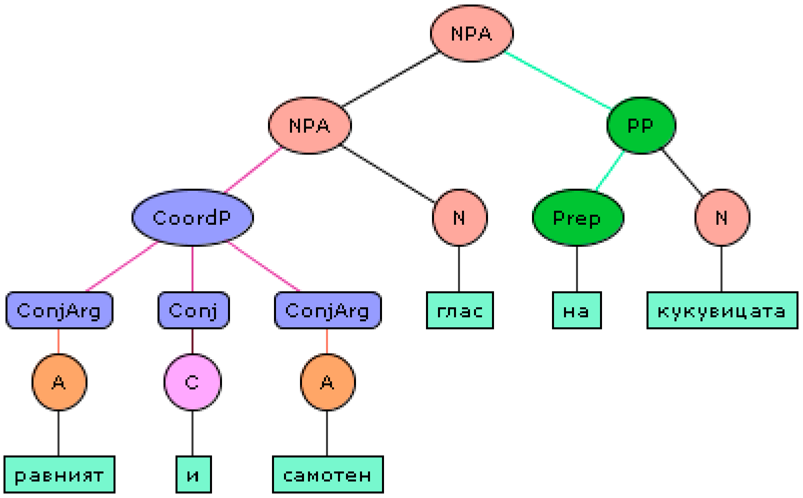
In coordinated sentences for purpose, when one of them is headed by 'da'-form and the other - by 'za da', then the two clauses *CLDA* and *CLZADA* are coordinated. Here is an example:



NP Internal Coordination

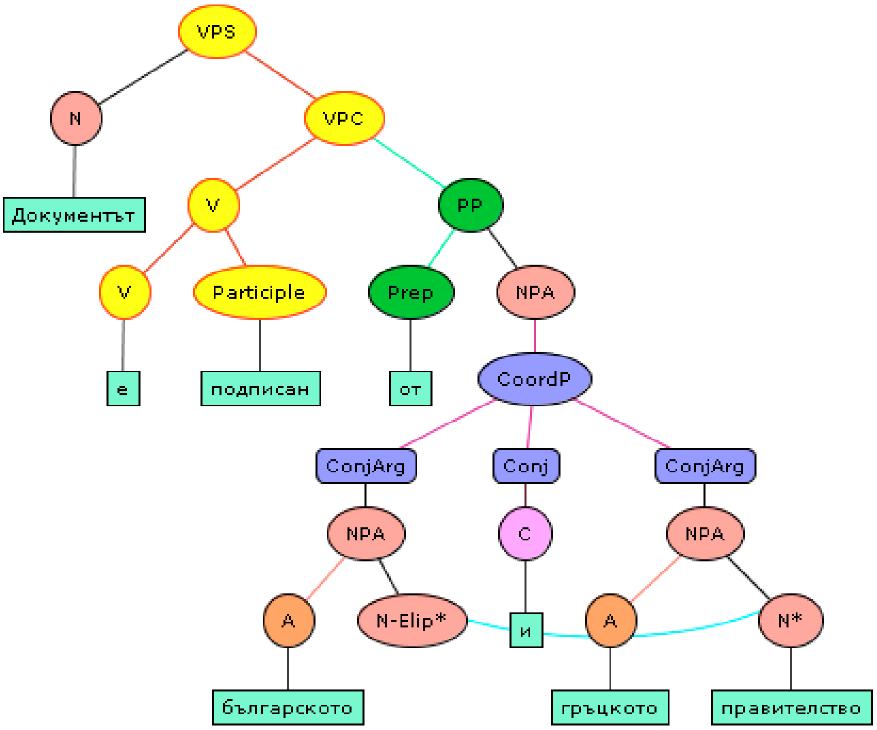
The *NP* internal coordination is a coordination of modifiers of the head noun (adjectives, adverbs, relative clauses, and prepositional phrases). Here is an example:

[равният и самотен] глас на кукувицата
[monotonic-the and lonely] voice of cuckoo-the
the monotonic and lonely voice of the cuckoo.



Here the coordination is between two adjectives, which are adjuncts (modifiers) of the head noun. They agree in their **MOD** feature. There are some cases in which this kind of coordination is not possible because the values of the **MOD** features do not agree. For instance:

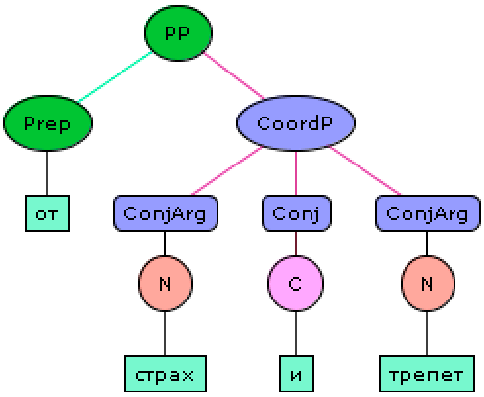
Документът е подписан от [българското и гръцкото] правителство.
Document-the is signed by Bulgarian-the and Greek-the government.
The Bulgarian and the Greek governments signed the document.



Here the potential coordination [българското и гръцкото] (the Bulgarian and the Greek) is not possible because the referents to which the semantics of the two adjectives is connected cannot be the same object. In cases like this we use ellipsis to distinguish the two referents. In this case the coordination is of two *NP*s, one of which is with an ellipsis (*N-Elip*). The blue curve shows the connection between the ellipted element and the noun.

NP coordination

This is a coordination of two saturated *NP*s. A crucial problem is the determination of agreement features of the coordination. They are set via principles similar to Head Feature Principle. Here we have coordination between two *NP*s, which are complements of the preposition.

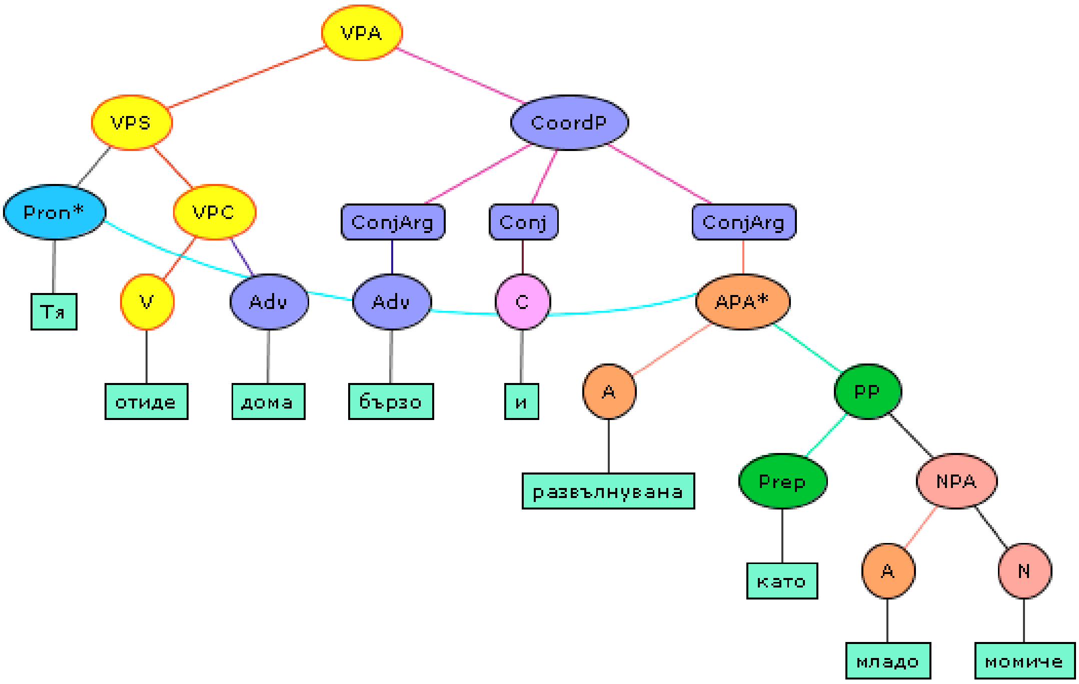


Adjunct Coordination

Sentential adjuncts can be coordinated when their **MOD** features agree in their values. The category of the coordination does not play any role in such constructions and can be underspecified without losing relevant information.

Тя	отиде	дома	[бързо	и	развълнувана	като	младо	момиче].
She	go[aor]	home	[quickly	and	excited	like	young	girl]

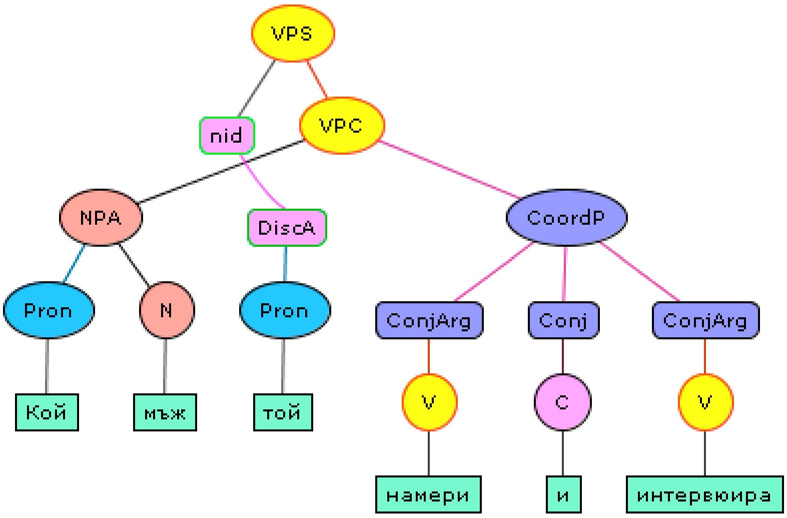
She went home quickly and excited like a young girl.



In this example the coordination phrase is an adjunct of the *VP* phrase. Their **MOD** features agree in their selection requirements being intersective adjuncts. The blue curve represents the fact that the second conjunct expresses secondary predication and it is co-referent with the subject 'тя' ('tja', she).

In the sentence below the two verbs are coordinated and they share the same object. Following the principle of dependents realization: complements -> subject -> adjuncts, the subject 'той' (toj, he) separates the verb-complement phrase and for that reason it is marked as *DiscA*, which means that it is realized as a constituent at a higher node:

Кой мъж той намери и интервюира?
Which man he found and interviewed?
Which man did he find and interview?

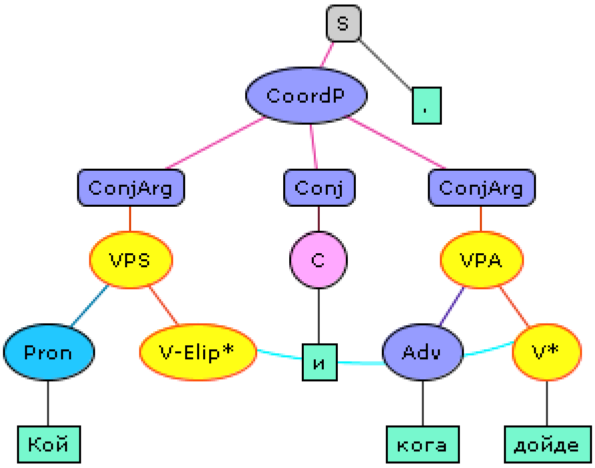


In all other non-discontinuous cases we treat subjects at sentential level with a co-reference mechanism to the *pro-ss* in the second clause. In cases of long-distance dependency the appropriate **SLASH** feature ensures the right head dependent realization.

The selectional preference of the head sometimes can be violated due to some specific properties of wh-words. For example, it concerns the possibility for a coordination of a subject and an adjunct. In such cases the ATB-like explanation is blocked and the only reasonable solution would be the ellipsis of the verb in order to keep the dependent realization consistent:

Кой и кога дойде?
Who and when came?
Who came and when?

When an ellipsis of the verb is introduced, the coordination is transferred to the sentential level. Thus, instead of coordinating dependents with two different grammatical roles, we coordinate two clauses:

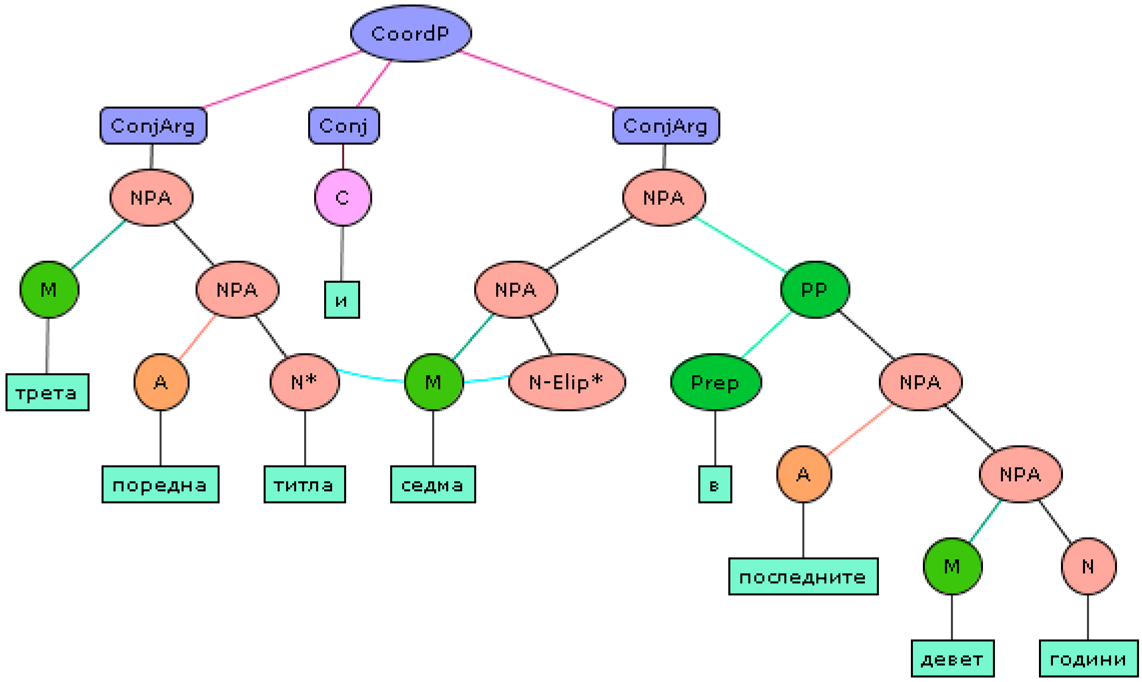


10 Ellipsis

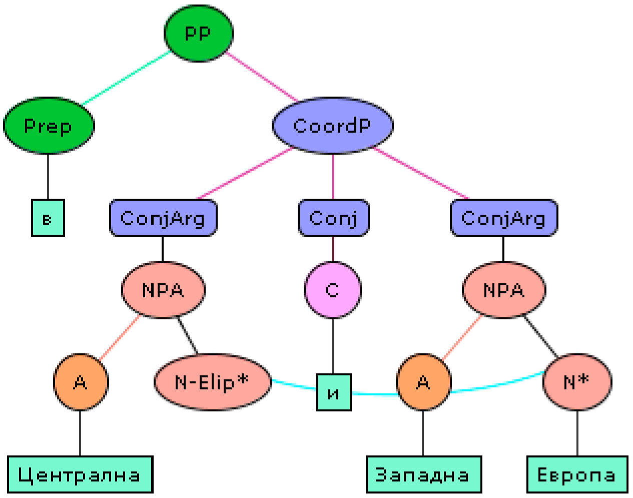
Elliptical material in the text is a frequent phenomenon involving mainly interference with coordination (subordination) or on discourse level. In our approach we introduce explicit elements for the missing material. These elements serve as anchors which show the constituent position of the missing element, and also they are connected to the elements that license the elliptical construction. Thus we consider the ellipsis as a slot in the elliptical phrase which determines what information is missing and where it can be copied from within the context. We assume that the word order of the missing material can be determined. This is a difference with respect to the extracted elements for which not always a sure position can be determined. Usually, the elliptical phrases are a missing head or a missing complement within a coordinated structure or in the discource structure.

We provide two types of analysis for ellipsis, depending on whether it is restored within the sentence or not:

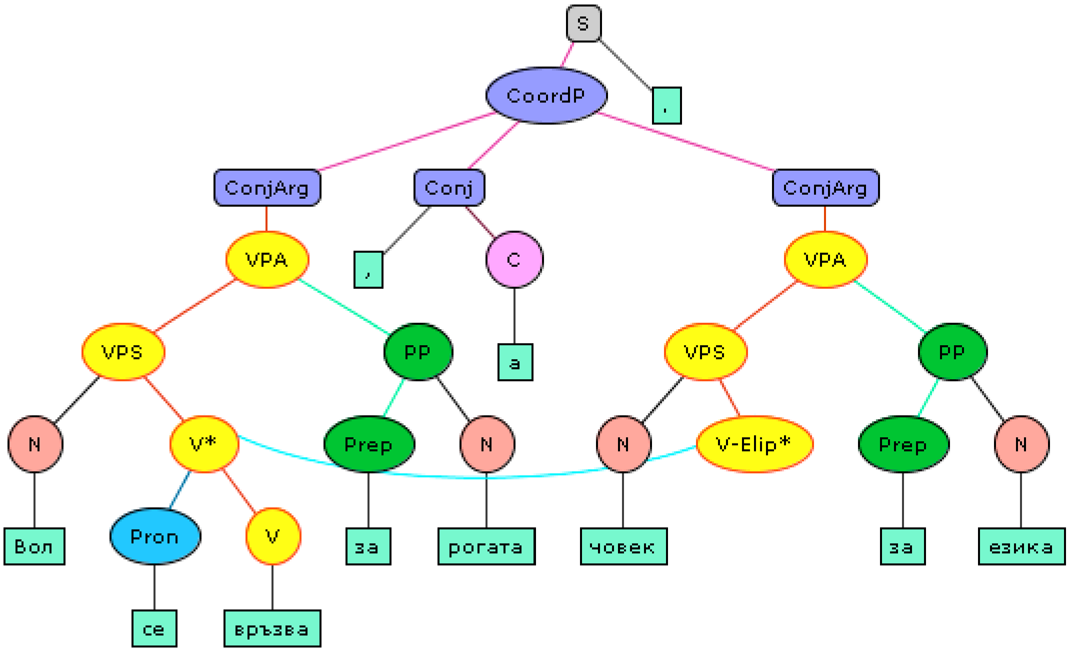
1. The ellipsis, which is recoverable within the sentence, is further governed by a reference mechanism, which establishes a connection with the overt material that supports the ellipsis. In the treebank we have four kinds of elliptical phrases. They are represented by the following elements: *V-Elip* (for verb or verb phrase), *N-Elip* (for noun or noun phrase), *Prep-Elip* (for a preposition) and *PP-Elip* (for a prepositional phrase). Note that the first two elements cover both - lexical and phrasal ellipsis. *V-Elip* has two attributes — *type* and *gram*, which indicate how the copied material is changed. The first attribute (*type*) shows whether the ellipsis equals the present form (*eq*), or it is a morphosyntactic variant of it (*var*), or it is the opposite (*neg*). The attribute *gram* is used when the ellipsis is a variant of the original verb or verb phrase to show in more exact way the variance of the characteristics. For example, a singular verb can be a trigger for a plural ellipsis. *N-Elip* has only a *gram* attribute which is reserved for the same purpose as in the previous case. *Prep-Elip* and *PP-Elip* do not have any attributes. We assume that each ellipsis points to the maximal phrase that was copied to it with some modification. The modification is specified by the *type* and *gram* attributes. Additionally, for the *V-Elip* it is allowed a *pro-ss* element to be attached. In this way it is possible the subject to participate in a co-referential relation. Here are some examples of sentence ellipses.



The above tree is an example of *N-Elip* element. The noun 'титла' ('titla', title) is copied to the second conjunct. As it was mentioned in the section on coordination, very often *N-Elip* is introduced in phrases of the following pattern: **A1 и A2 N**, where A1 and A2 stand for adjectives and N - for noun, but the two adjectives are such that their semantics cannot be united to point to one referent only. In such cases we introduce an *N-Elip* element. Here we give one more example of this case within the expression: в Централна и Западна Европа (in Central and West Europe):

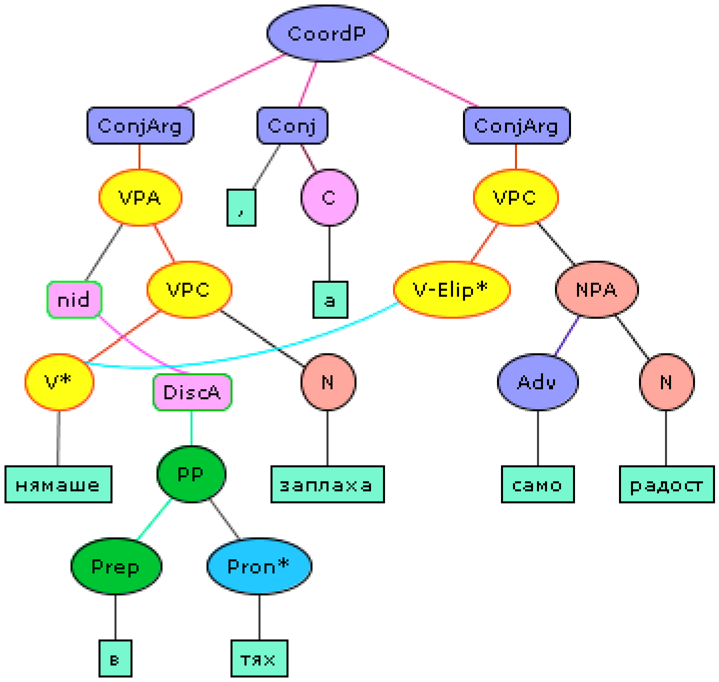


The following example demonstrates a *V-Elip*. In this case the explicit verb is copied without any change:



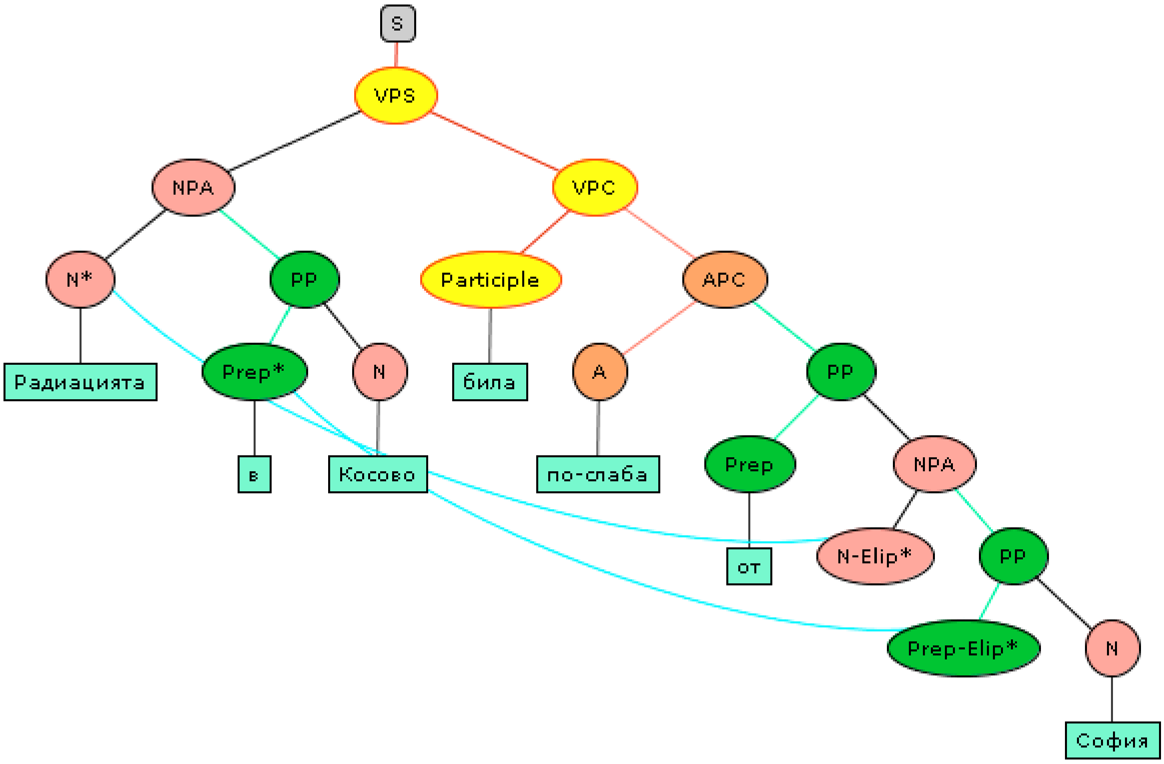
The next example contains a *V-Elip* with variation, which is presented only in the XML representation. In this case the ellipsis is the most extreme variation, namely - the opposite form of the original:

> Нямаше в тях заплаха, а само радост
> Nyamashe v tyah zaplaha, a samo radost
> There was not threat in them, but only joy

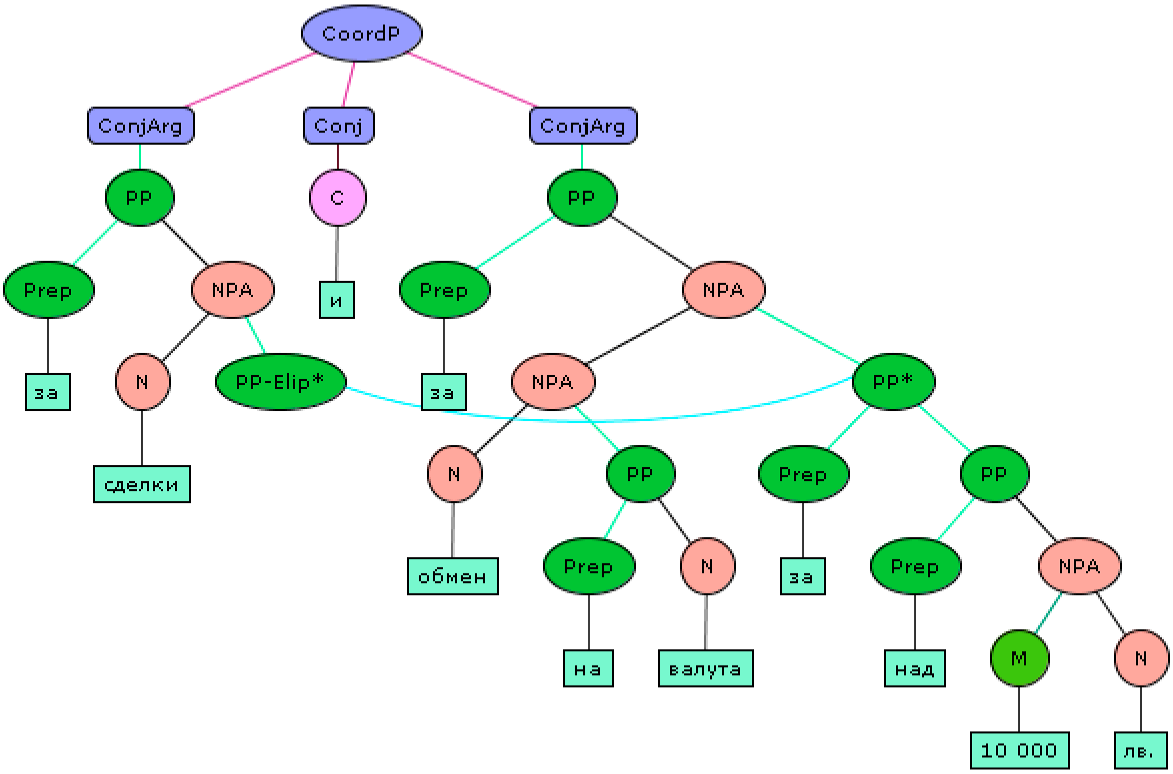


The V-Elip element expresses the opposite meaning of the negative verb 'нямаше' (there was not), namely — the positive verb 'имаше' (there was).

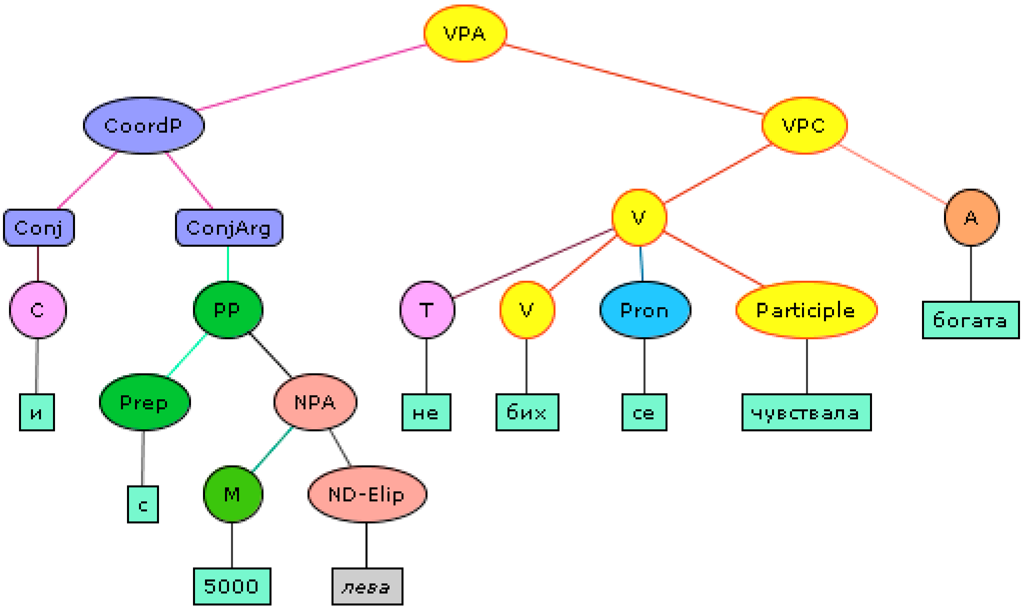
The next example demostrates an ellipsis of a preposition and of a noun.



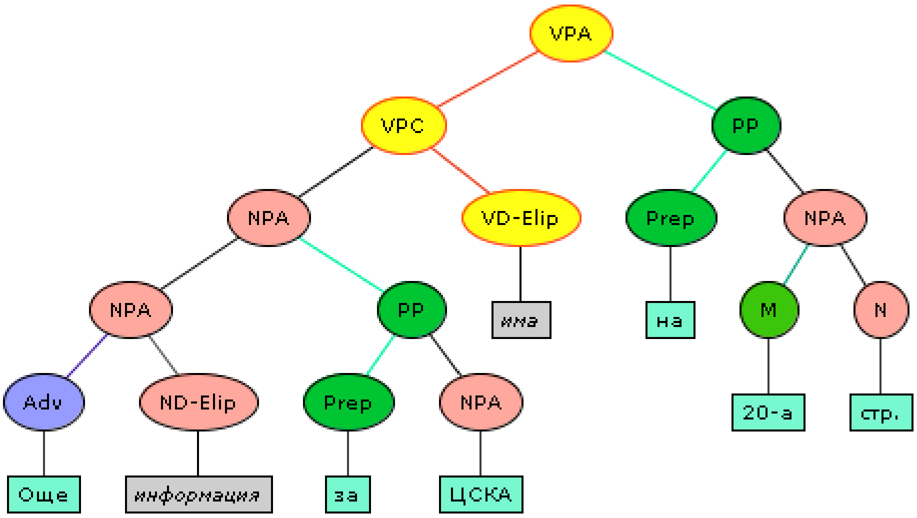
The last example here shows an ellipsis of a whole prepositional phrase. It is not possible to have a coordination of the two noun phrases instead of the ellipsis here because the noun phrases are complements of prepositional phrases.



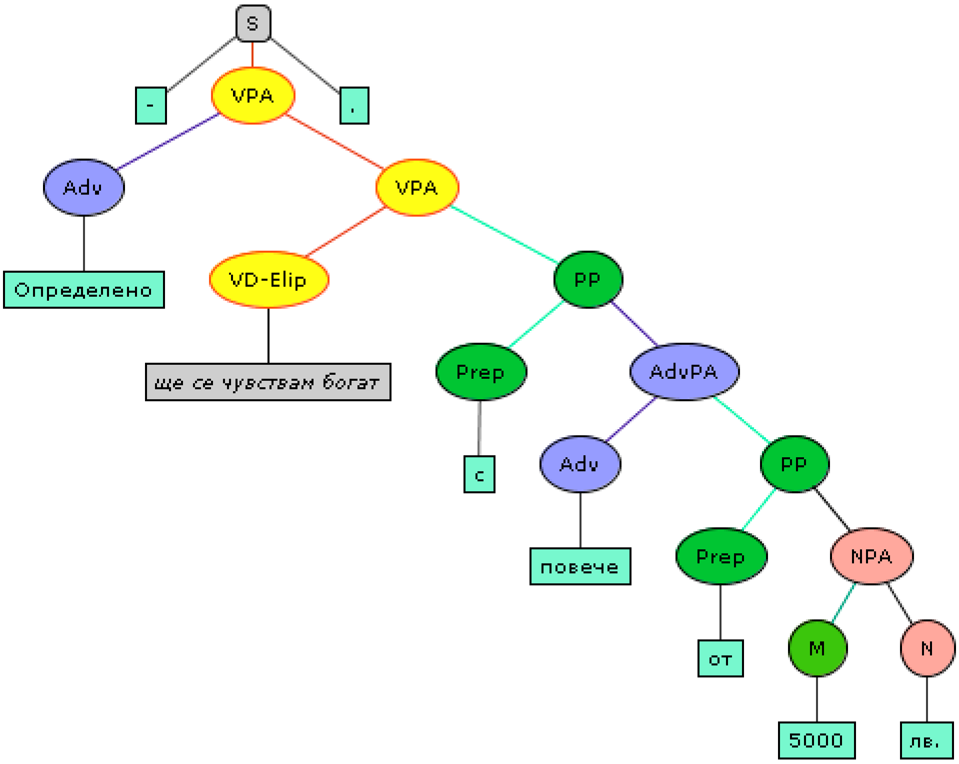
2. The non-recoverable in the sentence ellipsis is treated as a discourse one. We have three elements for this kind of ellipsis: *VD-Elip*, *ND-Elip* and *PPD-Elip*. All of these elements have three attributes: *type*, *gram*, *form*. The attribute *type* has two possible values: *world knowledge* and *discourse*. The value *world knowledge* means that the ellipsis can be recovered on the basis of our world knowdge only. The value *discourse* means that the ellipsis can be recovered on the basis of the discourse — in some of the neighbouring context. The element *VD-Elip* has an additional value: *exists* for the frequent ellipsis of the copula verb. The attribute *gram* is used to represent the morphosyntactic features of the ellipsis. The attribute *form* is used to represent the basic form of te missing material. Here are a few examples of such ellipses.



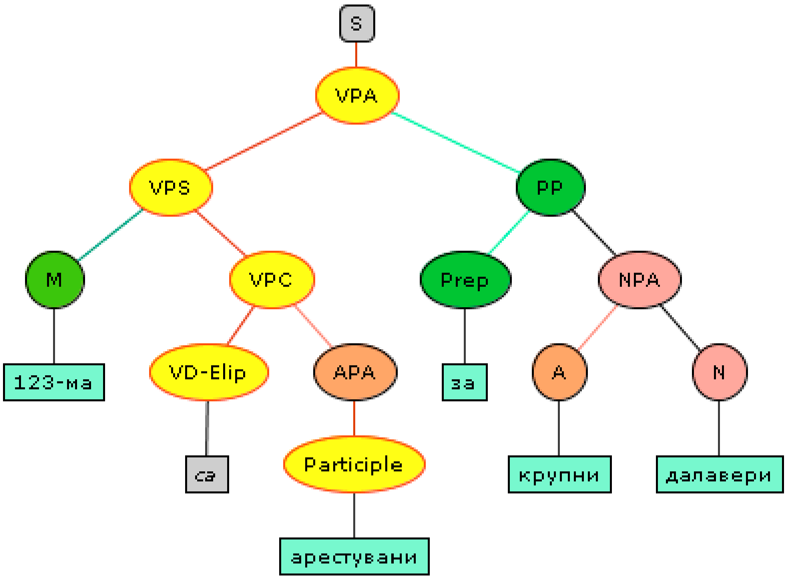
The example is of a *ND-Elip* which is determined by the discourse. The value of the form attribute is shown under the *ND-Elip* node. The next example shows *ND-Elip* recovered by our world knowledge and *VD-Elip* of type 'exists' with a form 'there is'. In this case we recover the ellipsis with specifying that *информация* (information) *има* (exists) on the corresponding page 20.



The next example demostrates the recovering of a whole phrase from the discourse:



Here is a typical case of a *VD-Elip*: in newspaper headlines, when the copula is systematically omitted:



11 Other Phrases

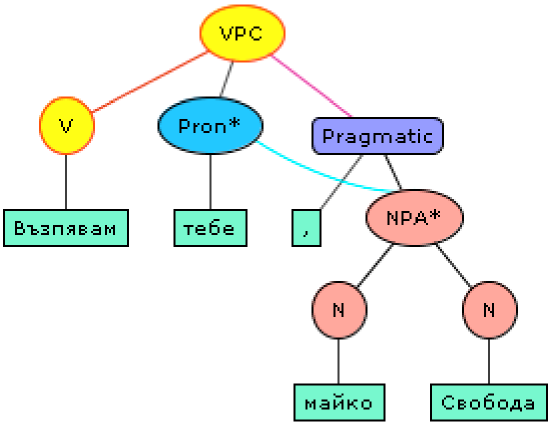
Here we present two types of phrases which have a special treatment: *pragmatic elements* and *foculizers.* They share the following property: they can appear together with (almost) all other types of phrases.

11.1 Pragmatic

Under pragmatic phrases we understand phrases that have mainly pragmatic contribution to the sentence analysis. Generally, we analyze pragmatic elements as adjuncts within a domain (nominal, verbal, etc). We do not introduce a separate constituent structure for the Pragmatic elements, but they are realized among the daughters of the constituent they modify. The most frequent Pragmatic elements are the vocative phrases, the evaluative adverbs with respect to a proposition, parenthetical expressions. We have an HPSG analysis of the vocative - see [Osenova and Simov 2002]. However, the other types still need more elaborate investigation.

Vocative Expressions

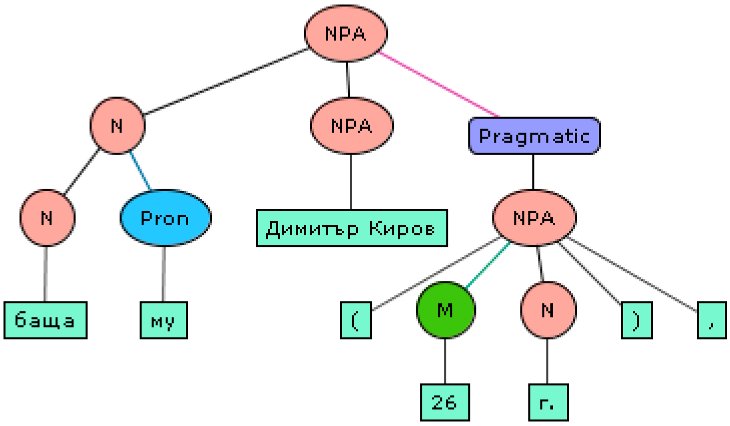
Vocatives are always attached to the sentence level inspite of their position within the sentence. They interact with the Context Indices in the HPSG representation (referring to the addressee of the utterance) and with its help sometimes they co-refer with other elements in the sentence. We explicate such co-referent relations, if present. Here is an example:



In the example the vocative phrase is attached to the *VPC* element, but we assume that it modifies the pragmatic features of the whole sentence. The vocative phrase is co-referent with the second person accusative pronoun, which is the complement of the VPC phrase.

Parenthetical Expressitions

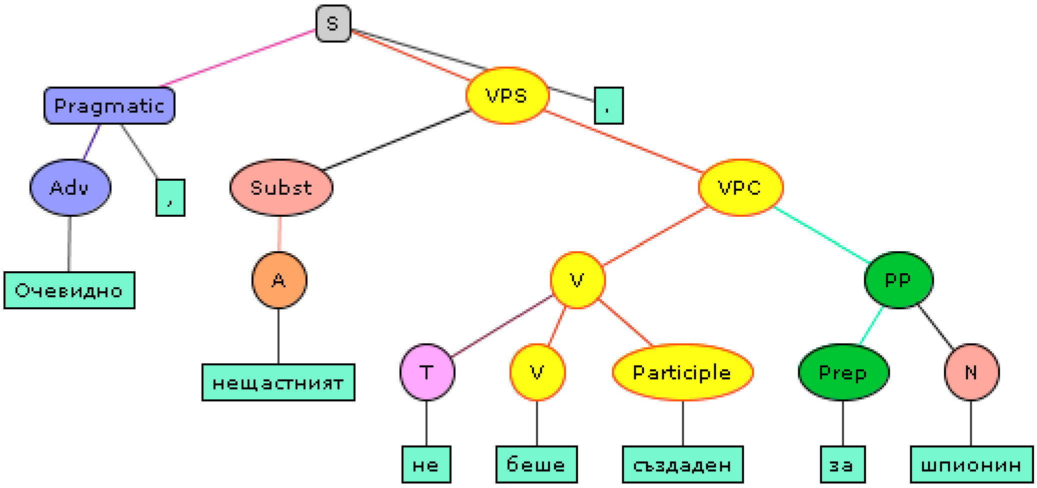
There are several kinds of these expressions, but we will present them in brief. Usually, they represent additional information to the pragmatic content of a constituent, or they express the attitude of the speaker to the utterance. Here is an example:



Here the Pragmatic element adds optional information about the age of the person denoted by its child element *NPA*.

Evaluative Adverbs

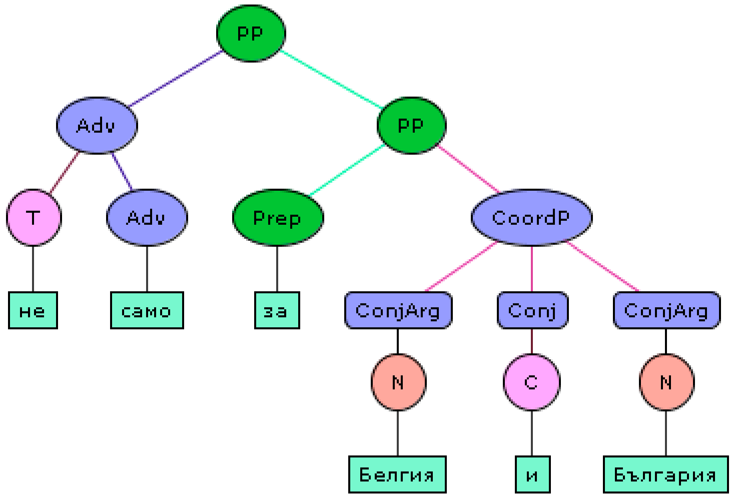
These are adverbs that express some modality with respect to the whole utterance. Here is an example:

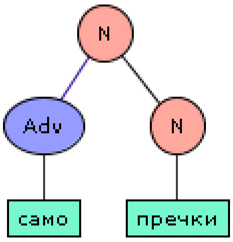


Here is a list of some of the most frequent adverbs of this kind: 'сигурно' ('sigurno', sure), 'очевидно' ('ochevidno', obviously), 'може би' ('mozhe bi', maybe), 'едва ли' ('edva li', hardly), etc.

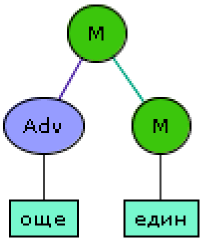
11.2 Foculizers

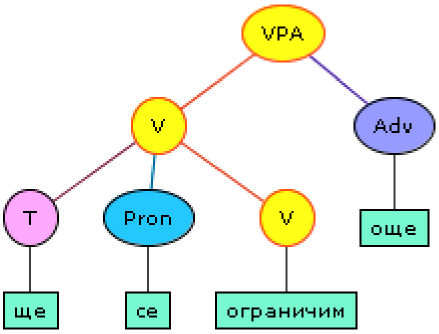
Here we include emphasizing words, which do not change the category and project the same phrase: *NPA* to *NPA*, *PP* to *PP* and so on. The typical one is 'само' ('samo' only). It is attached to the appropriate phrase and the name of the mother node is the same as the phrase's one (*PP*, *AdvP*, *N*, *NPA* etc). Other foculizer words are: 'поне' ('pone', at least), 'все' ('vse', always), 'най-вече' ('naj-veche', mainly), 'точно' ('tochno', namely), 'дори' ('dori', even), 'лично' ('lichno', personally), in certain cases. Here are two examples:





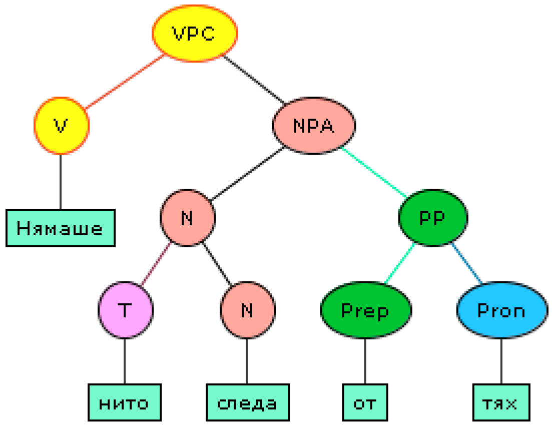
There are other words that can be used as foculizers, but at the same time they could project phrases: 'още' ('oshte'), 'чак' ('chak'), 'едва' ('edva'). In the following examples, first we present a case in which 'още' is a foculizer and then a case in which it is an adjunct.

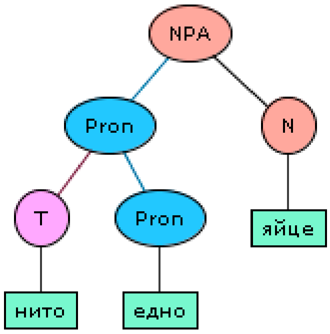




Sometimes the conjunction 'и' is treated as an emphasizing word. In the treebank we treat such cases as a coordinated phrase with one conjunct. In some contexts there are ambiguities with respect to the scope of the coordination. But as a rule we prefer to present the broader scope (see in Preference rules section).

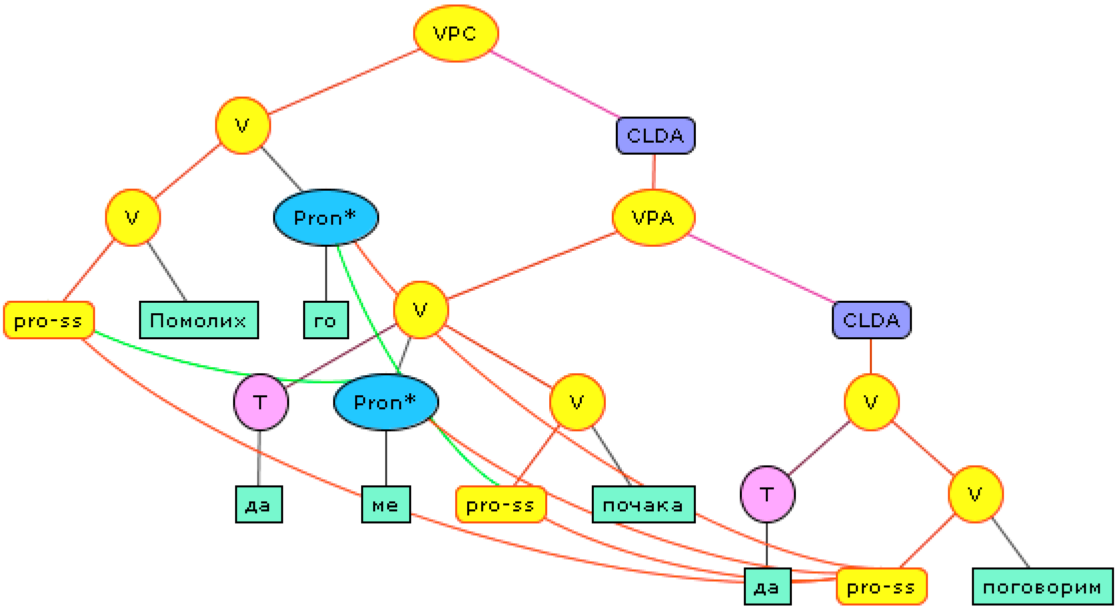
Another foculizer is the negative polarity items: 'нито' ('nito', nor). Here are examples:



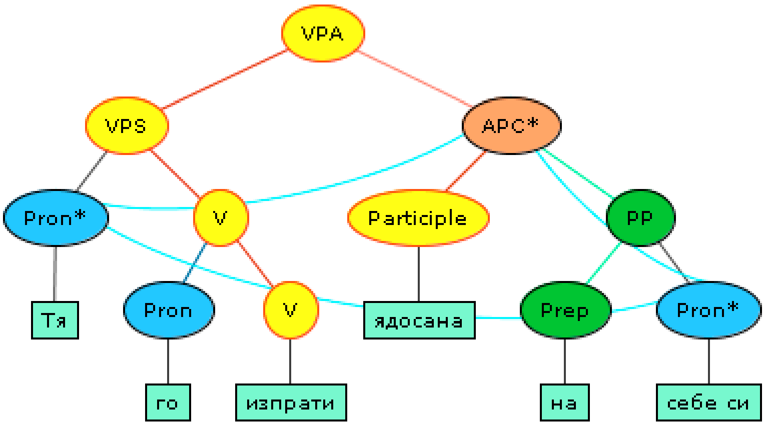


12 Co-reference

As it was already mentioned, we assume that each phrase (with some exceptions) is connected (directly or indirectly) to an index. Noun phrases and clauses introduce referent indices, while the adjective and adverb phrases add information about indices. The prepositional phrases do both. We consider the sharing (co-reference) of indices as crucial for the correct sentence analysis. In the treebank we explicate all the co-reference relations that cannot be inferred from the syntactic structure. We assume that the indices can refer to objects or sets of objects. Based on this assumtion we have encoded three relations: equality of objects (or sets); member-of and subset-of relations. Here is a list of language phenomena we explicate in the treebank using these relations: Anaphoric relations within the sentence, including the unexpressed subject of various kinds of clauses. Possession relation expressed by possesive pronouns. Secondary predication - a co-reference between an adjunct and a verb argument. Nominalizations - a co-referent relation between a clause (proposition) and a pronoun, or a noun (phrase). Cohesion chains - repetitions of the same noun phrase or different noun phrases with the same head. Some changed referring expressions like first and second person pronouns. Here we give some examples:



In this example (I have asked him to wait for me to talk) we have several co-referential relations: (1) equality of the unexpressed subject of the verb in the main clause with the accusative clitic of the verb in the first *CLDA* clause; (2) equality of the unexpressed subject of the verb in the first *CLDA* clause and the accusative clitic of the verb in the main clause; (3) member-of relations between the referents of the unexpressed subjects of the verbs in the main clause and the first *CLDA* clause and the unexpressed subject of the second *CLDA* clause. It means that the set of ‘we’ includes two members in this case: ‘me’ and ‘him’. Having in mind the properties of equality (symmetry) there are member-of relations expressed from the clitics to the unexpressed subject of the second *CLDA* clause.



In this example we have an equality relation between a secondary predication adjunct and an appropriate argument of the verb — the subject in this case. Also there is an equality relation between the reflexive pronoun and the index of the adjunct, and then - with the index of the subject of the sentence.

13 Preference Principles of Annotation

The basic assumptions in the annotation scheme are not suffice, because: there are always some mixed categories, which in different contexts behave differently, and for some phenomena there are more than one linguistically motivated possibility for an analysis. Thus some 'preference' rules have to be added to the basic desiderata. These principles are applied in cases, in which there exist at least two possible and linguistically motivated decisions. Thus the consistency of the annotation is guaranteed.

Here we list some cases, in which their role is of great importance. Not surprisingly, the 'preference' approach is suitable for well-known problematic phenomena like coordination and ellipsis. For example, we have formulated the following principles: *Prefer sentential coordination to pre-coordination* and *Prefer constituent coordination to ellipsis*. We take the first rule, because pre-coordination fits only in cases, when the selectional requirements of the verbs coincide. If the selectional preferences of two or more heads coincide, then the dependents are pre-coordinated, if not - then the preference rule is triggered. The preference of the sentential coordination guarantees consistency in all contexts. The second rule aims at decreasing the ellipsis in cases, when coordinated elements are of same dependence relation to the head.

Another dimension of preference rules application is the treatment of ellipsis. We distinguish between ellipsis, which can be restored within the sentence and ellipsis, which can be restored in the discourse or from our world knowledge. Sometimes the two interpretations are plausible, but the preference rule says: *If in the sentence there is an anchoring element for the ellipsis restoration, prefer it to the discourse one*. Thus some of the basic rules are:

1. Prefer sentential coordination to precoordination!(in cases when the scheme: complements, subject, adjuncts is violated, or when the valence requirements are different!). In all other cases prefer precoordination.

2. Prefer constituent coordination to ellipsis!

3. Prefer the broader scope of the coordinative conjunction и ('i', and) in a sentence initial position to the narrow one. The latter reading is not excluded, but only sometimes is made explicit.

4. In sentences where there are possible two readings of the modal verb: personal and impersonal, prefer the personal reading in order to avoid the subject extraction. The latter reading is not excluded, but only sometimes is made explicit.

In all other cases, in constructions with *modal verbs+da-forms*, where the subjects are identical, we follow the rules below:

- When the modal is personal, then the subject is realized according to its position - before the 'da'-construction it is the subject of this 'да'-construction; before the modal verb, it is the subject of the modal verb, respectively.
- When the modal is impersonal, then the subject, irrespectively of its position, is the subject of 'da'-construction. Here we include verbs like: може ('it is possible'), трябва ('must'), няма ('there is not') etc.
- In passive constructions, when the agent-like participant is expressed as a PP, consider it a complement of the verb.

5. In case of two-PP adjacent temporal or space expressions, prefer the mother PP analysis. If the two PPs are separated, then realize them one by one.

6. Concerning the scope of modal verbs over adjuncts, we follow the principle: If there is no change in the meaning, then the adjunct is attached to the most adjacent node.

7. In cases of unclear PP attachment in VP, we attach the PP as a nominal modifier when the connection is considered very tight. Otherwise, PP remains a second complement of the verb.

8. Within NPs there are the following preference rules of attachment:

- In NPA cases first the prepositive modifiers (adjectives, numerals, determiners) are realized, then postpositive (adverbs, *CLR*, *CLDA*, *CLQ*).
- In cases of type: един от X ('one of X') prefer substantivization to *N-Elip* analysis. Note that the phrase някой от X ('some of X') does not require substantivization because някой is an adjective and a noun in the lexicon.
- The family names are realized before the other modifiers.
- The noun identifiers of the names first realize their own modifiers and then are attached to the names.
- In NPC cases first the complement NP is realized and then, the prepositive modifiers of the head.

14 Conclusion

In this stylebook we described the architecture of the syntactic annotation of BulTreeBank. The analyses were performed with accordance to the HPSG language model, which ensured the presentation of constituent and dependency information, and an adequate interrelation between the syntactic, semantic and discourse levels. Both linguistic domains were discussed in detail — the lexical domain and the phrasal domain. Additionally, pragmatic elements and foculizers were introduced. The domains were presented in close connection with the linguistic phenomena like: coordination, subordination, ellipsis, complementation, co-reference etc. For the treatment of the complex cases we introduce a set of preference rules. The sentences were presented in a tree graphical view, which was designed in a user-friendly manner. With its rich linguistic information, the Bulgarian treebank can be used further for various applications like the development of a syntactic parser for Bulgarian, in E-learning education, in QA and IR tasks.

15 Acknowledgments

We would like to thank the annotators of the treebank for their devoted work and for their enthusiastic participation in many useful discussions during the annotation process: Elisaveta Balabanova, Dimitar Doikoff, Iliana Garavalova, Sia Kolkovska, Magdelena Paunova. We would like to thank Alexander Simov, Milen

Kouylekov, Hristo Ganev, Ilko Grigorov, Krasimira Ivanova for the development of the CLaRK software tool, which supported and facilitated our work. Special thanks go to Erhard Hinrichs, Sandra Kübler, Frank Müller, Manfred Sailer, Frank Richter, Heike Telljohann, Tylman Ule from SfS for their valuable comments and help. We also thank all the staff in SfS for the friendly and fruitful scientific environment during the visits in Tübingen.

References

[Andreychin 1976] Ljubomir Andrejchin. 1976. *Zalogat v bulgarskata glagolna sistema*. In: Pashov and Nitsolova (eds.) Pomagalo po bulgarska morfologiya. Glagol. Sofia, Nauka i izkustvo, pp. 60-76. (In Bulgarian)

[Bulgarian Academy Grammar 1983]:1983. Volume 3, Syntax (In Bulgarian)

[Boyadjiev et. al. 1998]: Todor Boyadjiev, Ivan Kucarov, Iordan Penchev. 1998. *Contemporary Bulgarian Language — Encyclopaedia*. Part III: Syntax, (In Bulgarian).

[Brezinski 2001]:Stefan Brezinski. 2001. *Bulgarian Syntax*. Sofia (In Bulgarian)

[King 1989] Paul King. 1989. *A Logical Formalism for Head-Driven Phrase Structure Grammar*. Doctoral thesis. Department of Mathematics, University of Manchester, Manchester, England.

[King and Simov 1998] Paul King and Kiril Simov. 1998. *The Automatic Deduction of Classificatory Systems from Linguistic Theories*. (Revised version). Grammars, 1(2): 103-153. Kluwer Academic Publishers, The Netherlands. 1998.

[Osenova and Simov 2002] Petya Osenova and Kiril Simov. 2002. *Bulgarian Vocative within HPSG Framework*. In: Proc. of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG), Kyung Hee University, Seoul, South Korea. pages 94-100

[Pollard and Sag 1994] Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, Illinois, USA

[Simov 2001] Kiril Simov. 2001. *Grammar Extraction from an HPSG Corpus*. In: Proc. of the RANLP 2001 Conference. Tzigov Chark, Bulgaria.

[Simov et al. 2002] Kiril Simov, Milen Kouylekov, Alexander Simov. 2002. *Incremental Specialization of an HPSG-Based Annotation Scheme*. In: Proceedings of the Workshop on 'Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data', the LREC conference, Canary Islands, Spain.

[Simov 2002] Kiril Simov. 2002. *Grammar Extraction and Refinement from an HPSG Corpus*. In: Proc. of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics. Trento, Italy.

[Simov and Osenova 2004a] Kiril Simov and Petya Osenova. 2004. *BTB-TR02: BulTreeBank Text Corpus of Bulgarian:Content, Segmentation, Tokenization*. BulTreeBank Technical Report BTB-TR2.

[Simov and Osenova 2004b] Kiril Simov and Petya Osenova. 2004. *BTB-TR04: BulTreeBank Morphosyntactic Annotation of Bulgarian Texts*. BulTreeBank Technical Report BTB-TR4.

[Wasow, Bender, and Sag 2003] Ivan a. Sag, Thomas Wasow, and Emily Bender. 2003. *Syntactic Theory: A formal introduction*. 2nd Edition. Stanford: CSLI Publications.