Special Linguistic Phenomena in the Bulgarian HPSG-based Treebank (BulTreeBank)

Petya Osenova and Kiril Simov
BulTreeBank Project
Linguistic Modelling Laboratory, IPP, Bulgarian Academy of Sciences
http://www.BulTreeBank.org
Demo description

1 Introduction

Currently the BuTreeBank comprises 214 000 tokens, a little more than 15 000 sentences. Each token is annotated with morphosyntactic information. Additionally the Named Entities are annotated with ontological classes as person, organization, location, and other. Based on HPSG theory the annotation scheme defines a number of phrase types which reflect both - the constituent structure and the head-dependant relation. Thus we have phrase labels with the explication of the dependant types like VPC (verbal head complement phrase), VPS (verbal head subject phrase), VPA (verbal head adjunct phrase), NPA (nominal head adjunct phrase) etc. Behind the constituent structures and the head-dependant relations the treebank also represents phenomena like coordination, ellipsis, pro-dropness, word order, secondary predication, control – see (Simov and Osenova 2003). We will focus on some of them in this demo presentation. The treebank is encoded in XML.

2 HPSG language model and the annotation scheme

HPSG (Pollard and Sag 1994) 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 *structure-sharing*. Within our treebank we rely on the standard sort hierarchy of signs: the sort *sign* with subsorts *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 hierarchy of phrases that we assume is the following:

```
sign
    PHON : phonlist
    SYNSEM : synsem
  word
  phrase
      DTRS : dtrs
    headed-phrase
        head-complement
        head-subject
        head-adjunct
          head-sem-adjunct
          head-pragmatic-adjunct
        head-filler
    non-headed-phrase
```

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 pragmatic adjuncts are the vocative phrases in Bulgarian - see (Osenova and Simov 2002). The *head-filler* phrases account for the cases of unbounded dependency. The *non-headed-phrase* is used for dealing with coordination phrases.

The linearization of the constituents in HPSG is separated from the constituent structure and in this way the theory allows for different orders of the same constituent structure and discontinuous realization of the constituents. This separation ensures the representation of the grammatical relations within the constituent structure. 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 -> adjuncts*. The actual number and kind of dependents is determined by lexical elements within each phrase. There are two selectional mechanisms in HPSG: *valency principle* which ensures the selection of complements and subjects of the head and *modifier principle* which is responsible for the selection of the head by an adjunct. The two principles work over the constituent structure of the sign. The valency principle ensures that for each phrase of type head-complement, the complement (COMP) valency list of the head daughter is the same as the concatenation of the synsem values of the non-head daughters. Similarly for the head-subject phrase. A sign with empty valency lists is *saturated*. The modifier principle stipulates that for each head-adjunct phrase the value of the MOD feature of the non-head daughter is the same as the synsem value of the mother. The other important principle that we consider is Head Feature Principle which says that for each headed phrase the HEAD value is equal to the HEAD value of the head daughter.

In order to reflect the above informally stated HPSG theory we designed an annotation scheme which encodes the following information: *constituency* - each sign is represented as a phrasal label corresponding to the *category* of the sign; *head dependant relation* - the labels in the annotation scheme reflect the type of the phrase (head-complement, head-subject etc*);* *linear order* - the original word order is preserved and where necessary discontinuous elements are introduced; *co-referential relations* - each non-inferable co-referential relation is stated explicitly; *unexpressed elements* - unexpressed subject and ellipsis are represented explicitly. Pro-dropness is a characteristic feature of Bulgarian and when necessary it is represented as a *pro-ss* element. In each headed phrase the head daughter is represented implicitly and can be inferred automatically. The mechanism of *co-reference* is used for phenomena like pro-dropness, secondary predication, binding etc. Thus, in the annotation scheme (encoded as an XML DTD) the following types of elements have been distinguished:

Syntactico-phrasal elements

VPA(djunct) for verbal head-adjunct phrases, NPC(omplement), etc

Lexical elements

N, V, Prep, etc

Functional elements

Disc(ontinous), E(xtracted)

3 Facing some discourse-dependant phenomena

Here we focus on both – coordination and ellipsis, because they are connected to each other and because they proved to be the 'bottlenecks' in all linguistic theories.

3.1 Coordination

CoordP is an XML element which corresponds to a *non-headed* phrase. The coordination phrase in our analysis has a flat structure. It has only functional children which mark-up the role of each constituent: an argument of the coordination or a conjunction. A *ConjArg* XML element which represents an argument (conjunct) of a coordination phrase; a *Conj* XML element which marks-up a conjunction of a coordinated phrase together with the comma for the conjunction. 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

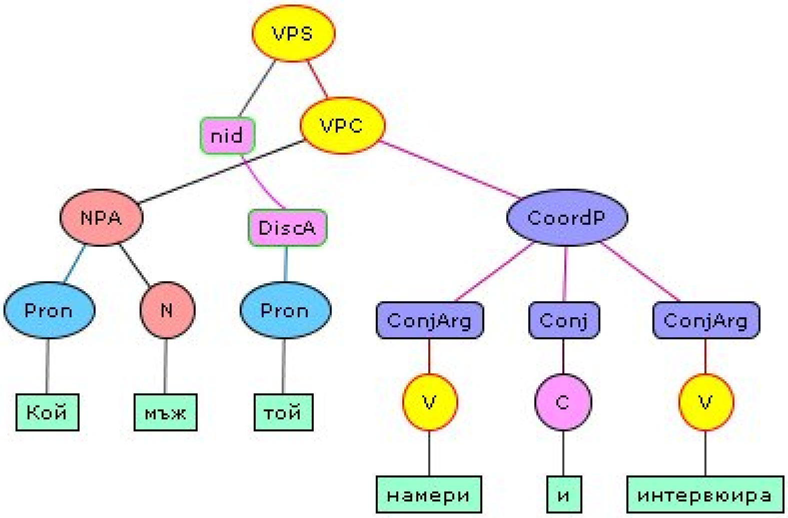
In this respect we propose not to classify obligatorily each coordination as an NP coordination, a VP coordination, etc., but to classify the coordination as *saturated coordination, adjunct coordination* and *unsaturated coordination.* Saturated coordination has empty valence lists, and the MOD feature 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.

We distinguish between the following types of coordination:

1. lexical coordination
2. clausal coordination
3. NP internal coordination
4. NP coordination
5. adjunct coordination

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 XML elemen *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?

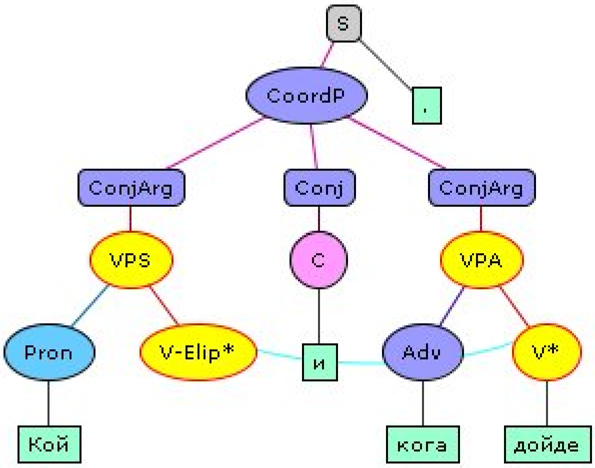


In all other non-discontinuous cases we treat subjects at sentential level with a co-reference mechanism to the *pro* in the second clause. In cases of long-distance dependency the appropriate SLASH feature ensures the right head dependant realization.

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 dependant 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:

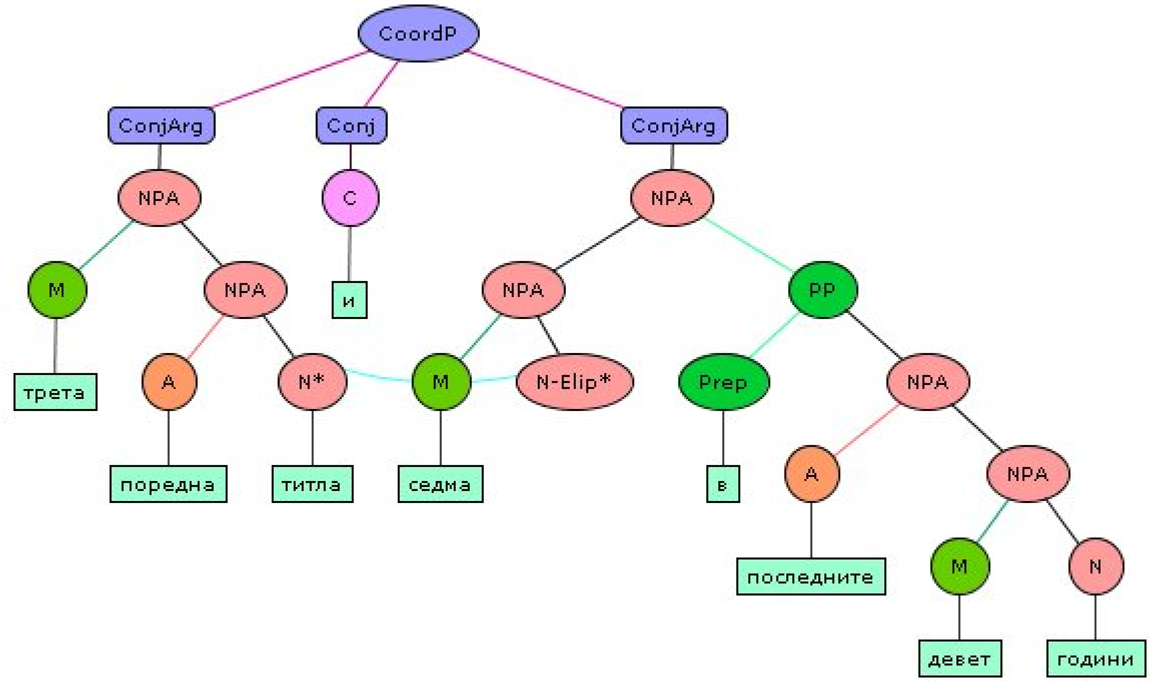


3.2 Ellipsis

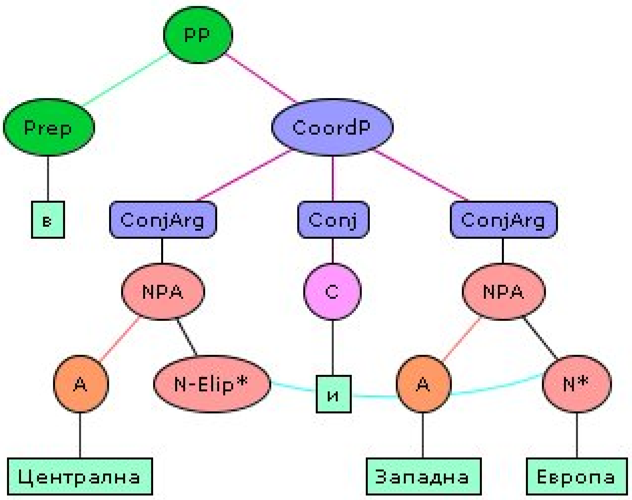
Elliptical material in text is a frequent phenomena involving mainly contrasts within coordination (subordination) on discourse level. In our approach we introduce elements for the missing material. These elements serve as anchors where the missing element has to be and also it is connected to the material which supports the elliptical construction. Thus we consider an ellipsis as a position in the elliptical phrase which determines what information and where is copied from the context. We assume that the word order of the missing material can be determined. This is different with respect to 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 larger context.

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

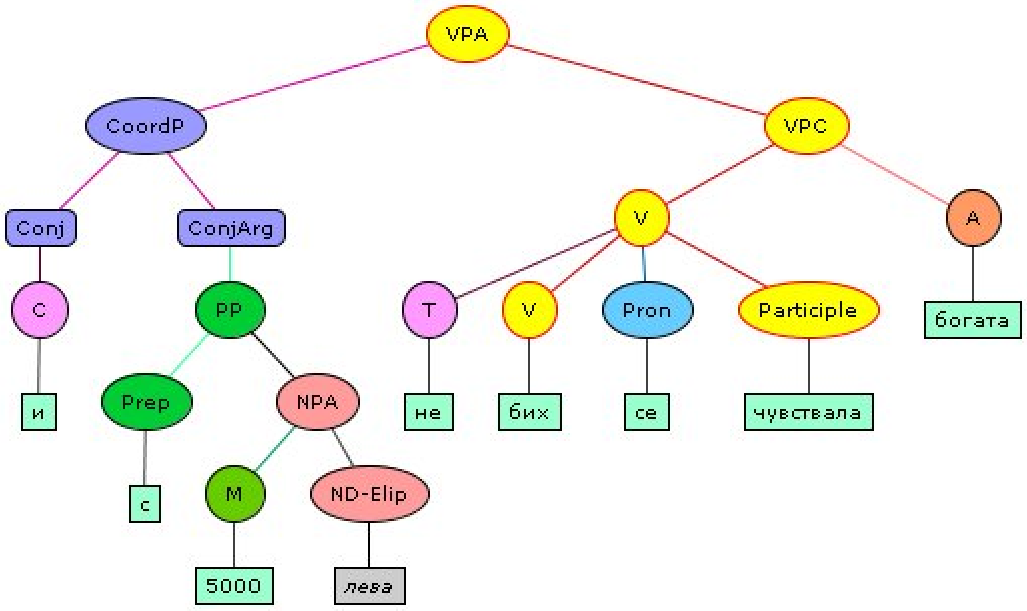
The ellipsis, which is recoverable within the sentence, is further governed by the following reference mechanism to the overt material that supports the ellipsis. In the treebank we have four kinds of elliptical phrases. They are represented by the following XML elements: *V-Elip* (for verb or verb phrase), *N-Elip* (for noun or noun phrase), *Prep-Elip* (for a proposition) and *PP-Elip* (for a prepositional phrase). The first two elements cover both - lexical and phrasal ellipsis. *V-Elip* has two XML attributes – *type* and *gram*, which indicate how the copied material is changed. The first attribute (*type*) shows whether the ellipsis equals the present from (*eq*), 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 what the new caracteristics are. For example, a singular verb can be a trigger for plural ellipsis. *N-Elip* has just *gram* attribute which serves the same purpose as in the prefious case. Prep-Elip and PP-Elip do not have any attributes. We assume that each ellipsis points to the maximal phrase that is copied 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 to change the subject of a verb phrase or to express it when it is not possible to copy it with the verb phrase. Here are some examples of sentence ellipses.



This is an example of *N-Elip* element. The noun 'титла' ('titla', title) is copied in 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** (A1 and A2 N), where A1 and A2 stand for adjectives and N for noun, but the two adjectives are such that their semantics can not be unionized for one referent. Thus in this case there is *N-Elip* element. Here we give an example of this case:



The non-recoverable in the sentence ellipsis is treated as discourse one. We have three XML elements for this kind of ellipsis: *VD-Elip*, *ND-Elip* and *PPD-Elip*. All of these elements have three XML attributes: *type*, *gram*, *form*. The attribute *type* has two values: *worldknowledge* and *discourse*. The value *worldknowledge* means that the ellipsis can be recovered on the basis of our world knowdge. The value *discourse* means that the ellipsis can be recovered on the basis of the discourse – some of the neighbouring sentences. 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 the missing material. Here is an example of such ellipsis from an interview. Here the elliptical part is "лева" (levs) and it is recoverable from the question at the beginning: Would you consider yourself rich in Bulgaria with 5000 levs? The answer in this sentence is: even with 5000 (levs) I would not feel rich.



4 Conclusion

Handling phenomena like coordination and ellipsis poses also the question of annotation consistency. For that reason we used “preference rules” in cases where more than one linguistic solution was plausible. For example:

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.

The encoding of other phenomena will be demonstrated as well. A detailed stylebook can be found on the web page: http://www.bultreebank.org/TechRep.html

References

[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 and Osenova 2003] Kiril Simov and Petya Osenova. 2003. *Practical Annotation Scheme for an HPSG Treebank of Bulgarian*. In: Proc. of the 4th Workshop on Linguistically Interpreteted Corpora, Budapest, Hungary.