The Language Model in Bulgarian Treebank (BulTreeBank)

Petya Osenova
(Sofia)

26.03.2007, Prague

Outline of the Talk

- General overview
- Synopsis of Annotation principles
- The levels of linguistic knowledge representation
- From HPSG to dependency
- Handling some linguistic phenomena
- Outlook and Conclusions

General overview: statistics

- HPSG-based format:
  - *Sentences*: 15 114
  - *Tokens*: 215 109
- Dependency format:
  - *Sentences*: 13 221
  - *Tokens*: 196 151

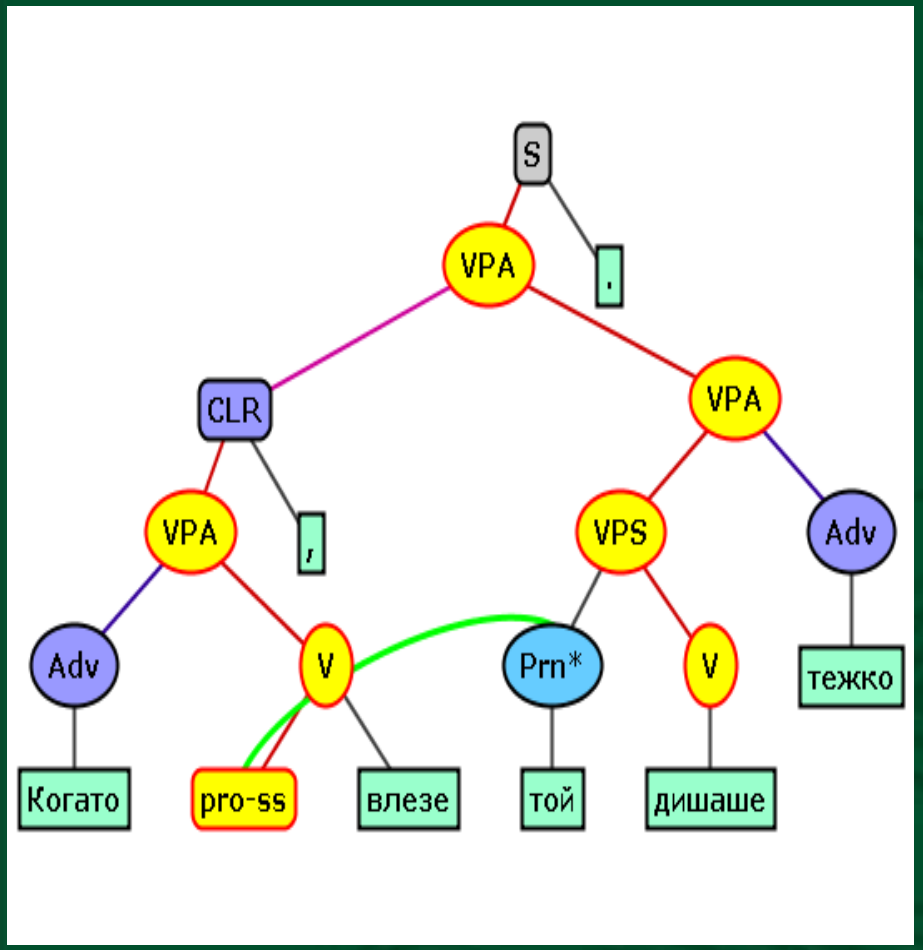
General overview: encoded linguistic information

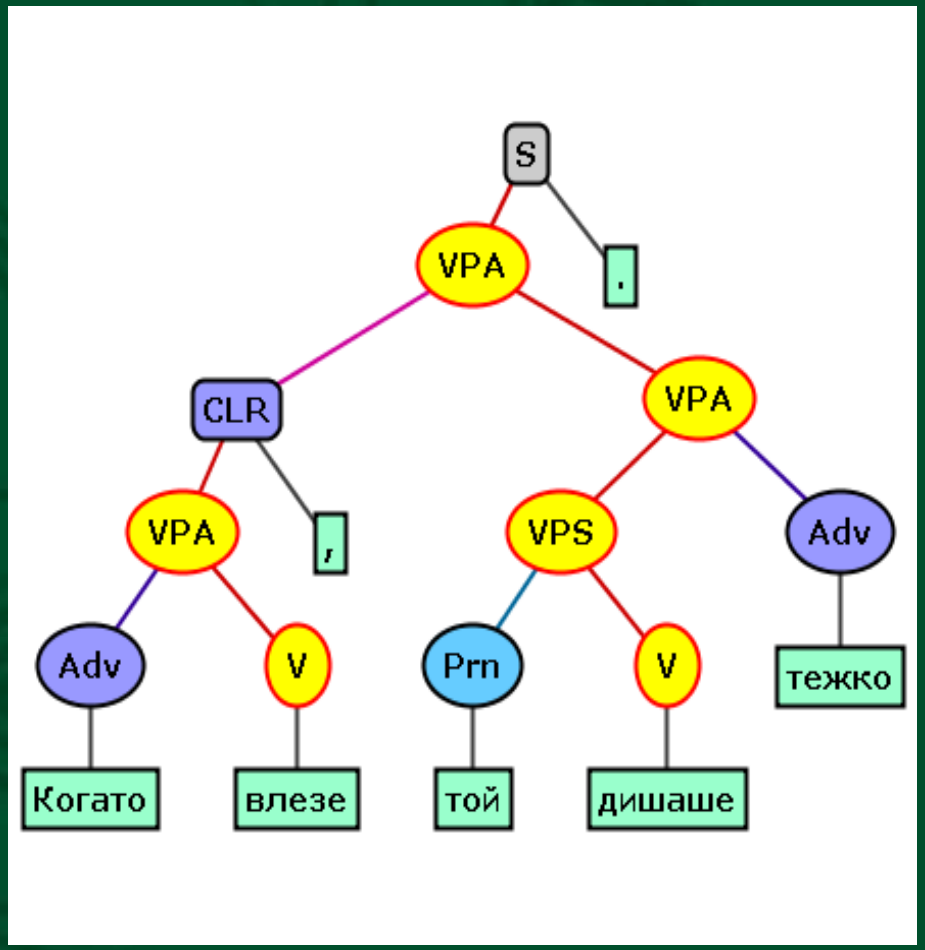
- Constituent structure (NP, VP, AP etc.) with crossing branches
- Dependency information (head-complement, head-adjunct, head-subject relations)
- Functional information (clauses, pragmatic elements, discontinuous elements)
- Encoded ellipsis and coreference relations
- Conversion to dependency structures exists (note that coreference and ellipsis are not presented in this format)

General overview: sources

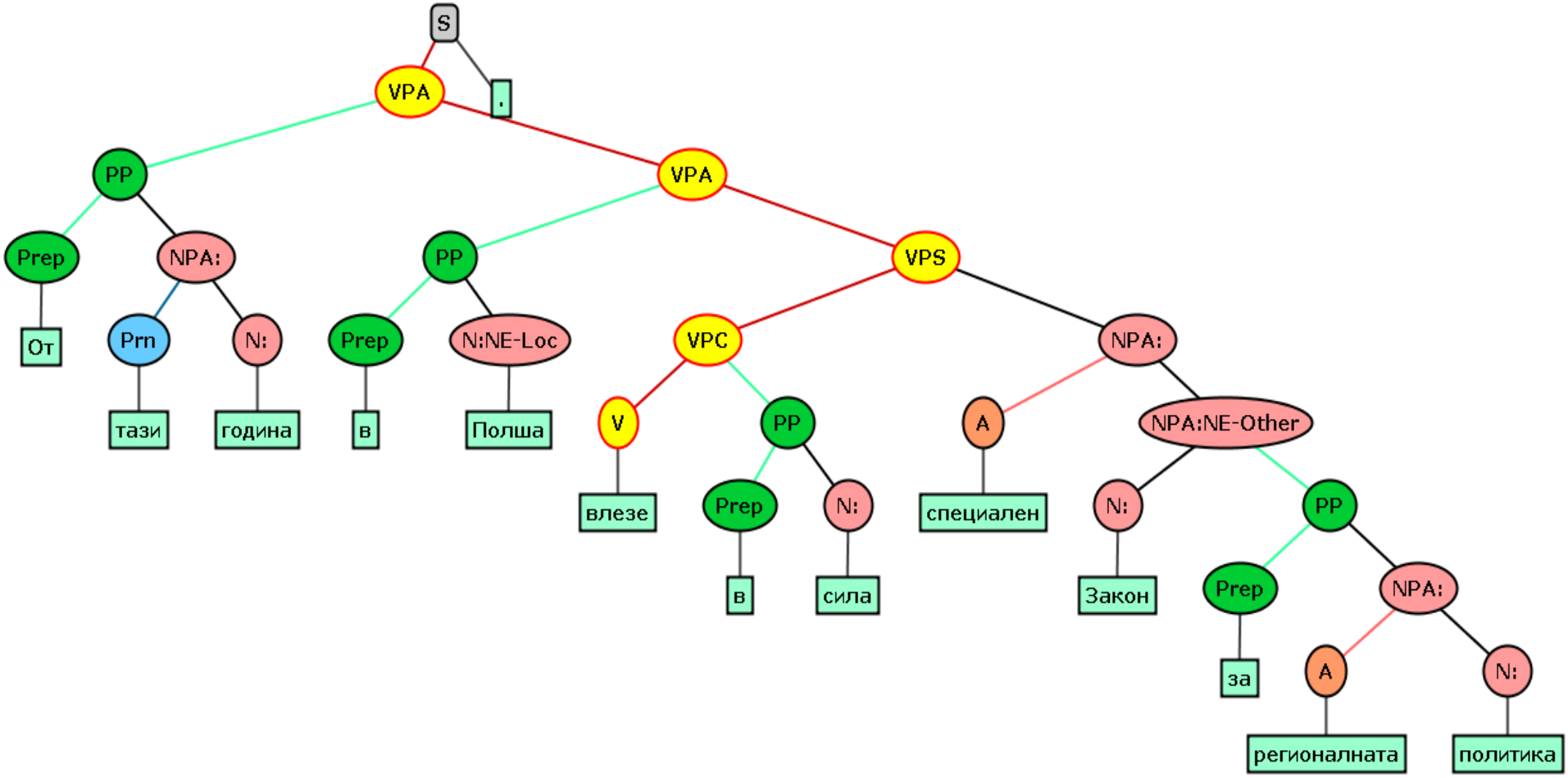
- Bulgarian grammars (1750 sentences)
- Randomly selected sentences (2027 sentences)
- Whole articles/texts from newspapers and literature (11 300 sentences)

Sentences from grammars: ambiguous interpretations





Sentence from random layer



Annotation Principles

- Theoretical aspects (HPSG language model)
- Implementational aspects (XML presentation)

HPSG-based Language Model: overview

- Linguistic objects
- Sort hierarchy (linguistic ontology)
  - Represents the main types of linguistic objects and their characteristics
- Grammar (theory)
  - HPSG Universal and Bulgarian Specific Principles
  - Bulgarian Lexicon

HPSG-based Language Model: Principles

- Head Feature Principle
- Valence Principle
- Adjunct Principle
- *Semantic Principle*

HPSG-based Language Model: hierarchy

- headed-phrase
  - head-complement
  - head-subject
  - head-adjunct
    - head-sem-adjunct
    - head-pragmatic-adjunct
  - head-filler
- non-headed-phrase

HPSG-based Language Model: constituency vs. dependancy

- HPSG separates the linear order from the constituent structure
- Each constituent structure reflects the dependency between its immediate constituents
- The realization of the dependants follows the sequence:

  complements > subject > adjuncts

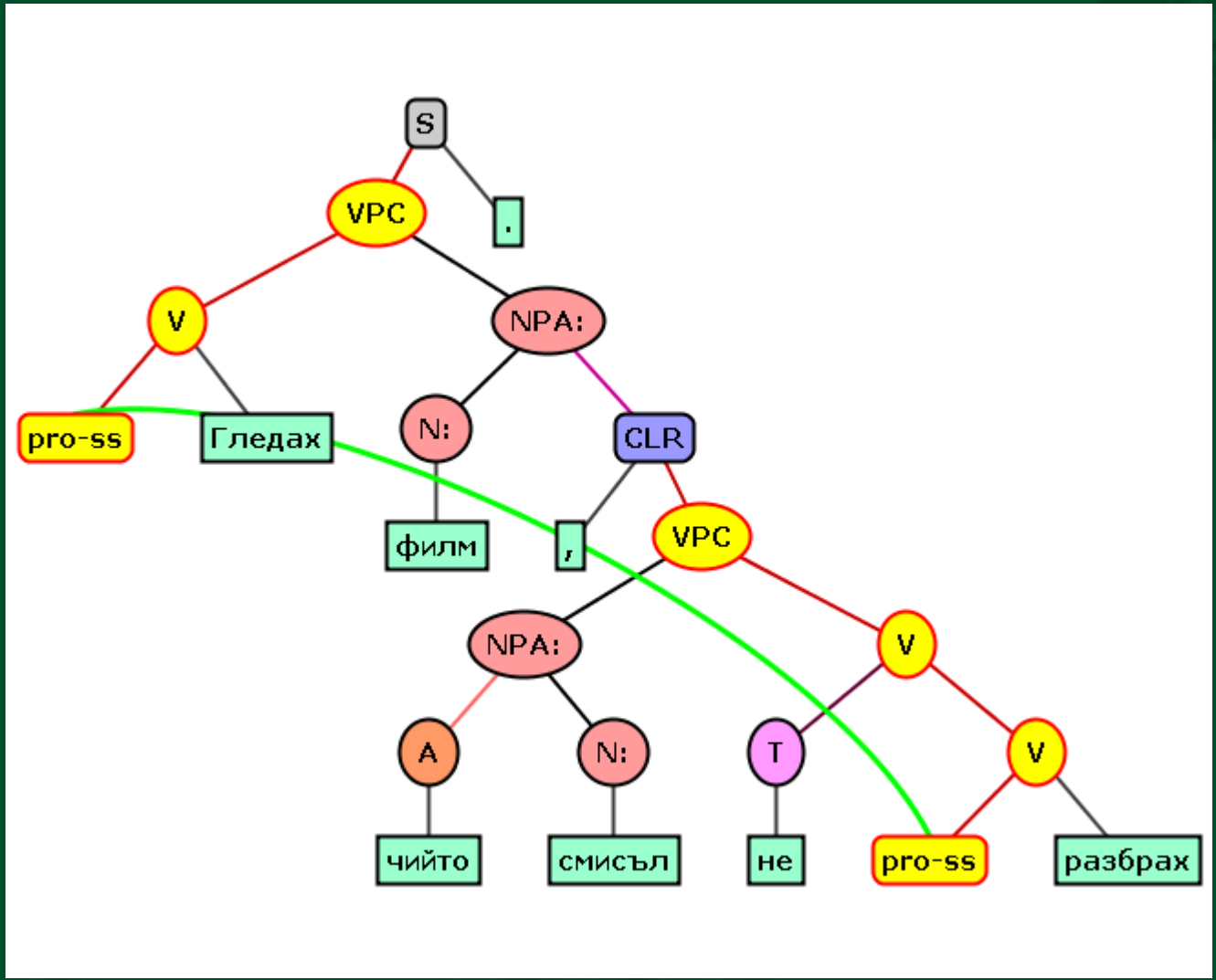
Implementation

- XML additive trees
- Graphical counterparts

The Levels of Linguistic Knowledge Representation

- Wordforms
- Morphological information
- Lexical elements (N, V, Prep)
- Syntactic elements (PP)
- Named Entities
- Dependency reflection (VPA(djunct), NPC(omplement)
- Functional elements (Disc(ontinous), Pragmatic)
- Relations – coreference, discontinuity, ellipsis

Example tree

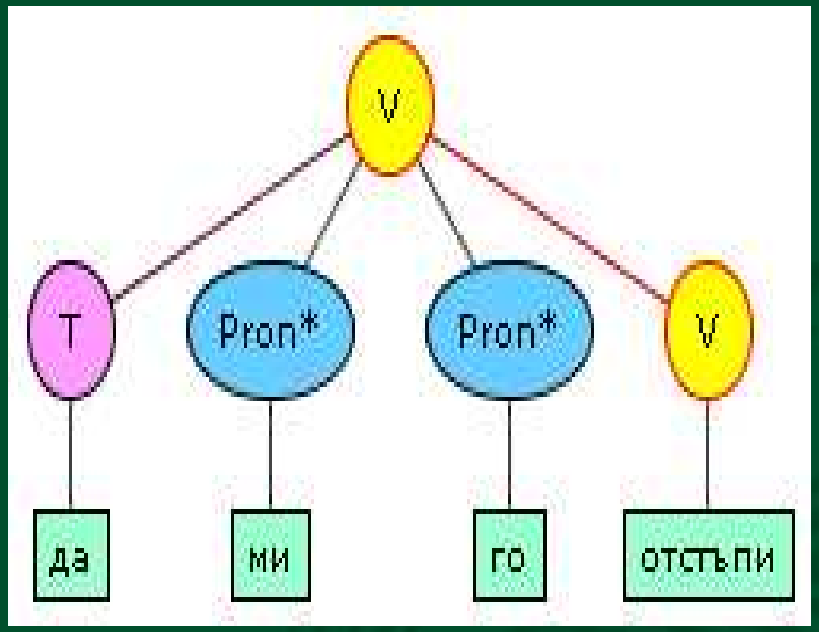


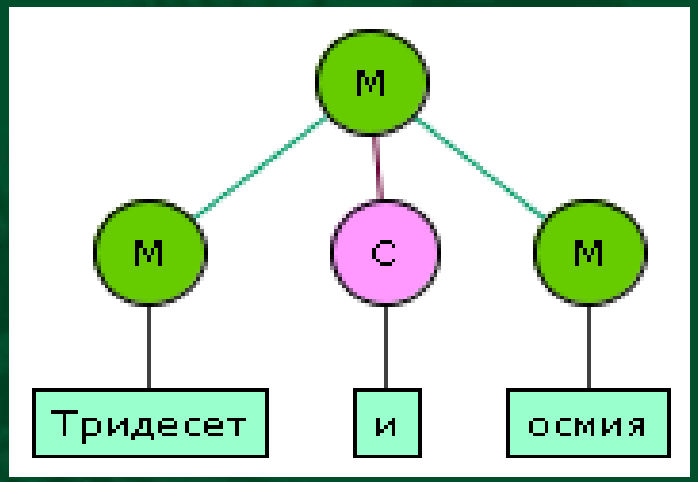
Lexical elements

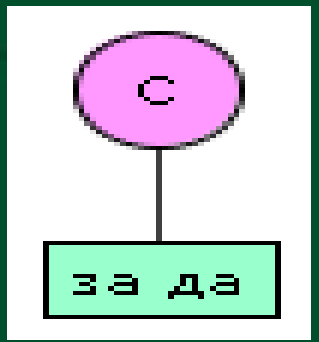
- Simple: one orthographic word
- Complex: verbal complex (head with clitics, da-construction), numerals, subordinators, prepositions

Foculizers: project lexical element, when combined with a lexical element

Examples of complex lexical signs







Syntactic elements

Traditional domains:

- NP
- VP
- AP
- AdvP
- PP

Functional elements

- Clauses – *CL, CLDA, CLCHE, CLZADA, CLR, CLQ*
- Sentence – *S*
- Markers of the coordination - *Conj, ConjArg*
- Markers of the elipsis – *V-Elip, N-Elip*
- Markers of discontinuity – *DiscA, DiscE, DiscM*
- Pragmatic elements - *Pragmatic*
- Non-immediate dominance - *nid*

Dependency reflection

- Explicit

NPA, NPC, VPC, VPA, VPS, PP, APA, APC, APA, AdvPA, AdvPC

- Non-explicit

CoordP, PP

The Dependency Representation

- No ellipsis presented
- Non-projective trees
- Two graphical views

A short description of the Dependency Part of BulTreeBank

This distribution represents only the dependency information encoded in BulTreeBank HPSG-based Treebank of Bulgarian

It contains sentences from Bulgarian Grammar Textbooks, Newspapers, Literature and other sources of texts.

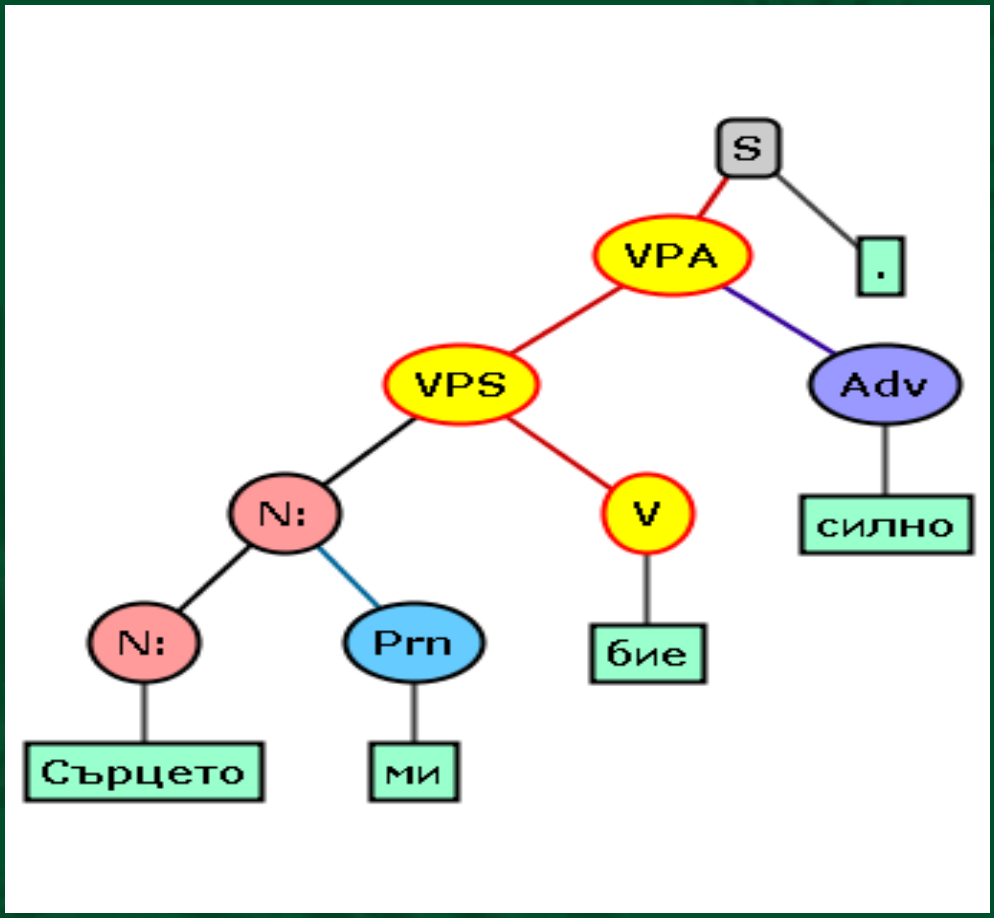
Full documentation (Style Book, Tagset description) of the Treebank can be found on: http://www.bultreebank.org/TechRep.html .

Dependancy links:

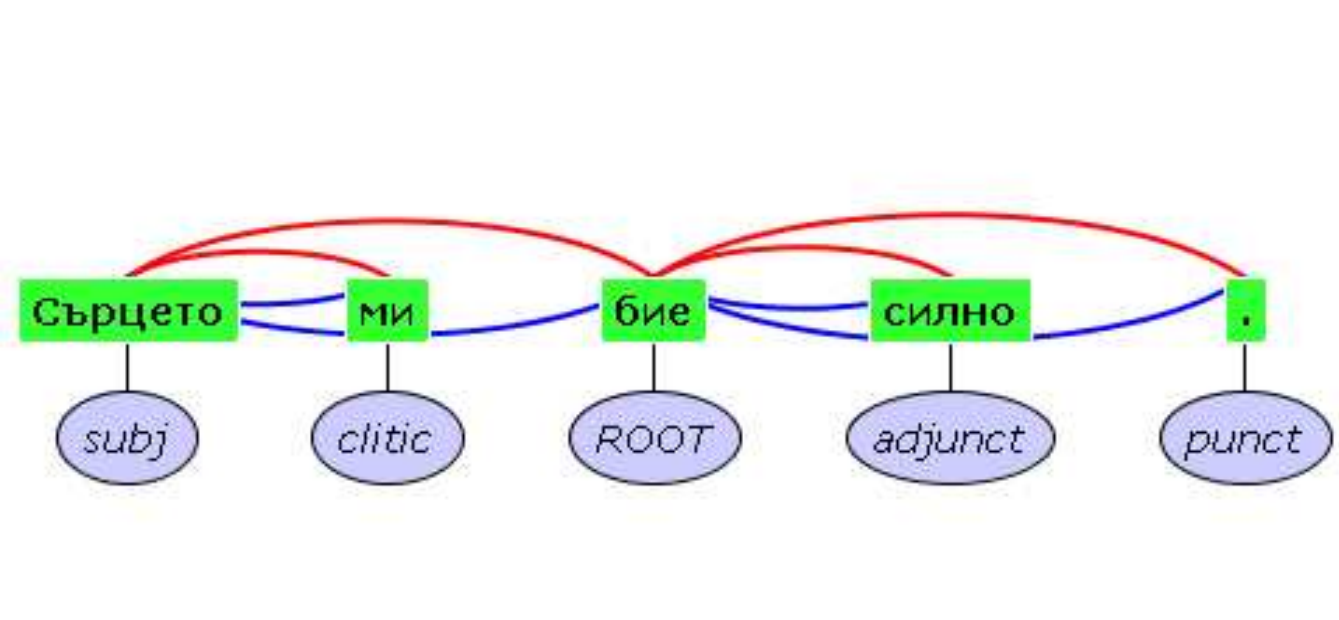
adjunct	Adjunct (optional verbal argument)
clitic	Clitic form
comp	Complement (arguments of: non-verbal heads, non-finite verbal heads, copula)
conj	Conjunction in coordination
conjarg	Argument (second, third, ...) of coordination
indobj	Indirect Object (indirect argument of a non-auxiliary verbal head)
marked	Marked (clauses, introduced by a subordinator)
mod	Modifier (dependants which modify nouns, adjectives, adverbs)
obj	Object (direct argument of a non-auxiliary verbal head)
pragadjunct	Pragmatic adjunct
prepcomp	Complement of preposition
punct	Punctuation
subj	Subject
xadjunct	Clausal adjunct
xsubj	Clausal subject
xmod	Clausal modifier
xcomp	Clausal complement
xprepcomp	Clausal complement of preposition

The conversion of the treebank was done by Kiril Simov, Petya Osenova, Svetoslav Marinov, Atanas Chanev.

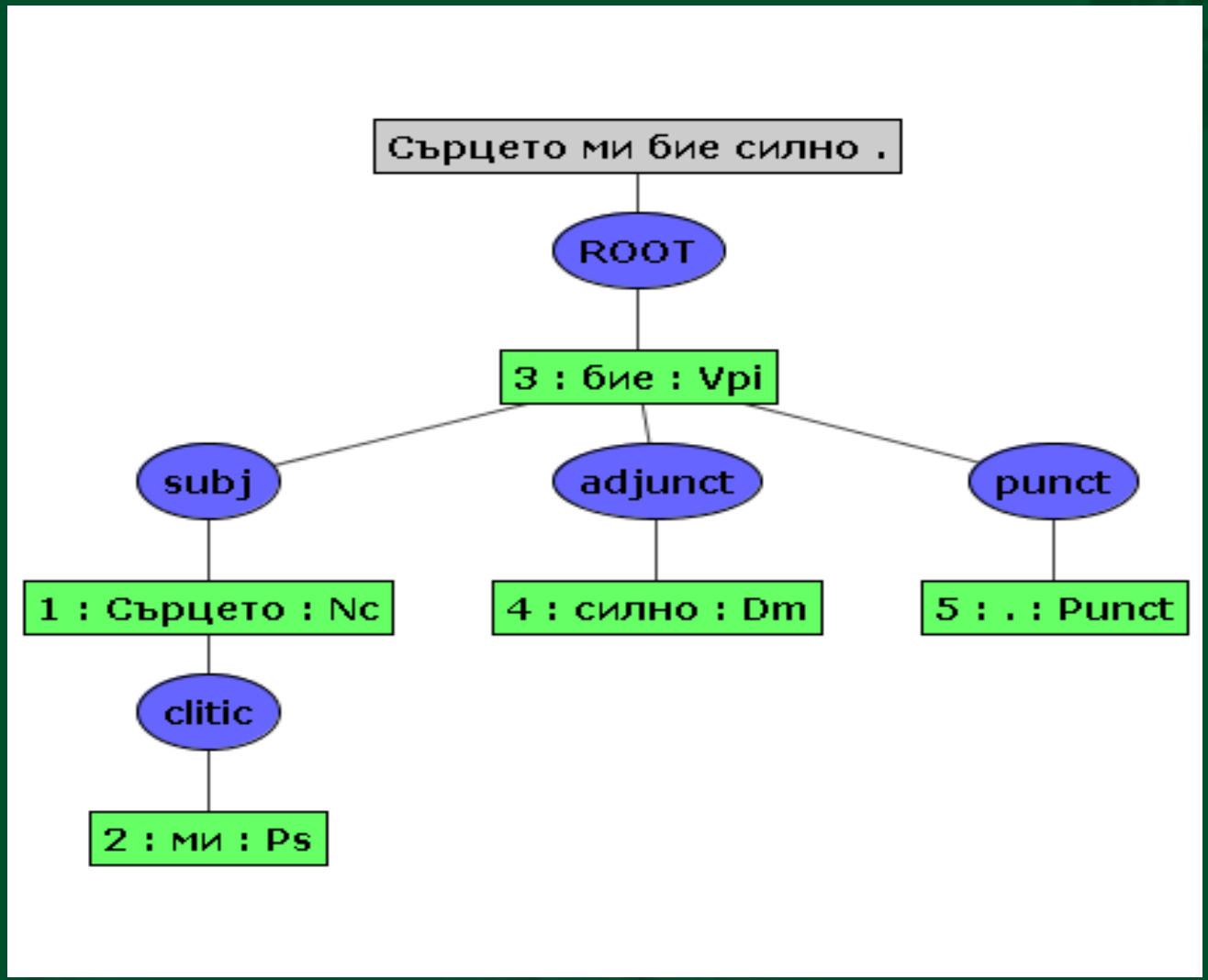
The original tree



The dependency-arrow presentation



Dependency-leaf tree



Handling some linguistic phenomena

- Word order
- Coordination
- Ellipsis
- Coreference
- Pragmatic expressions and foculizers

Word order

Two possibilities:

- permutations of elements that do not cause crossing branches effect
- discontinuity that violates the order of dependant realization

Permutations

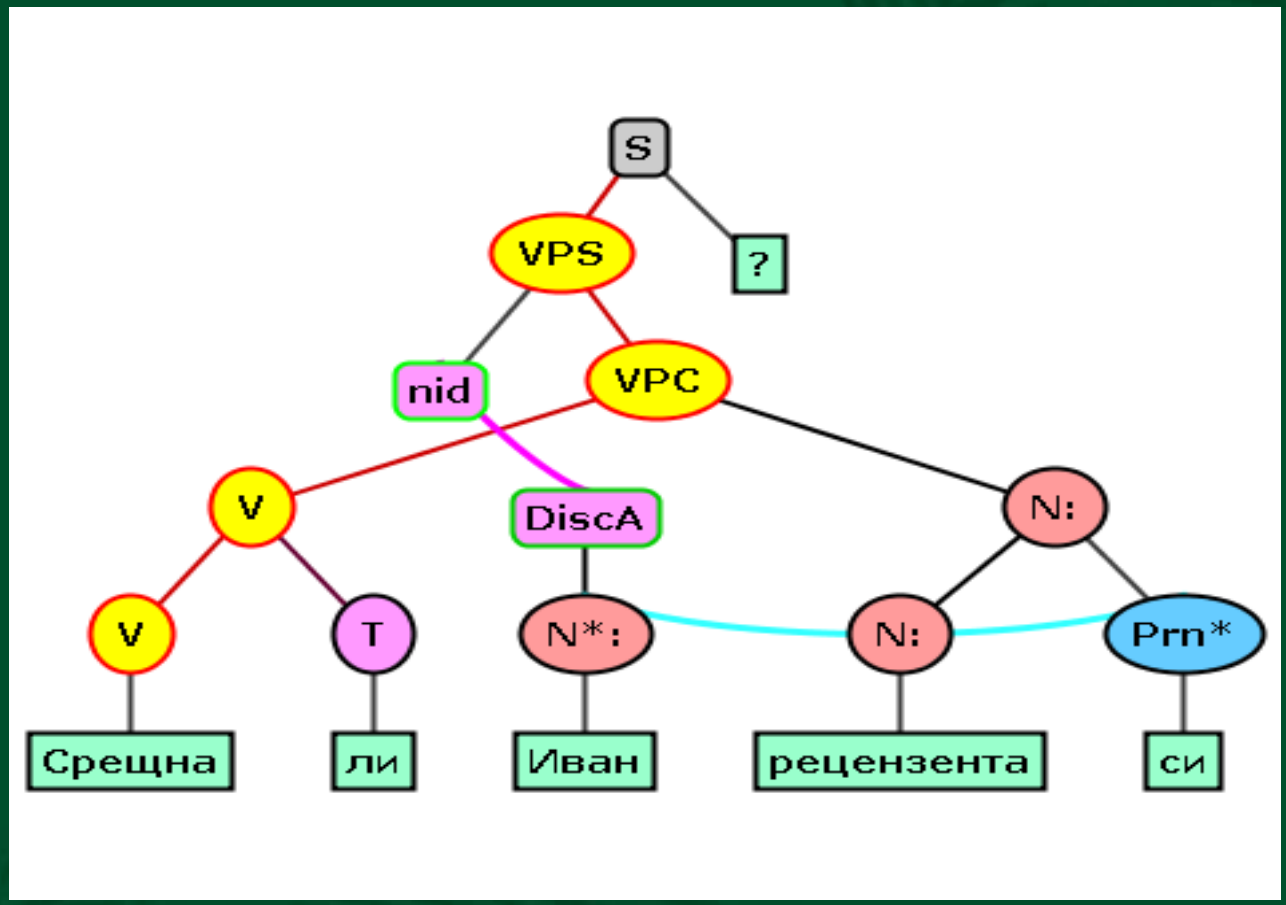
Петя посещава Прага

Посещава Прага Петя

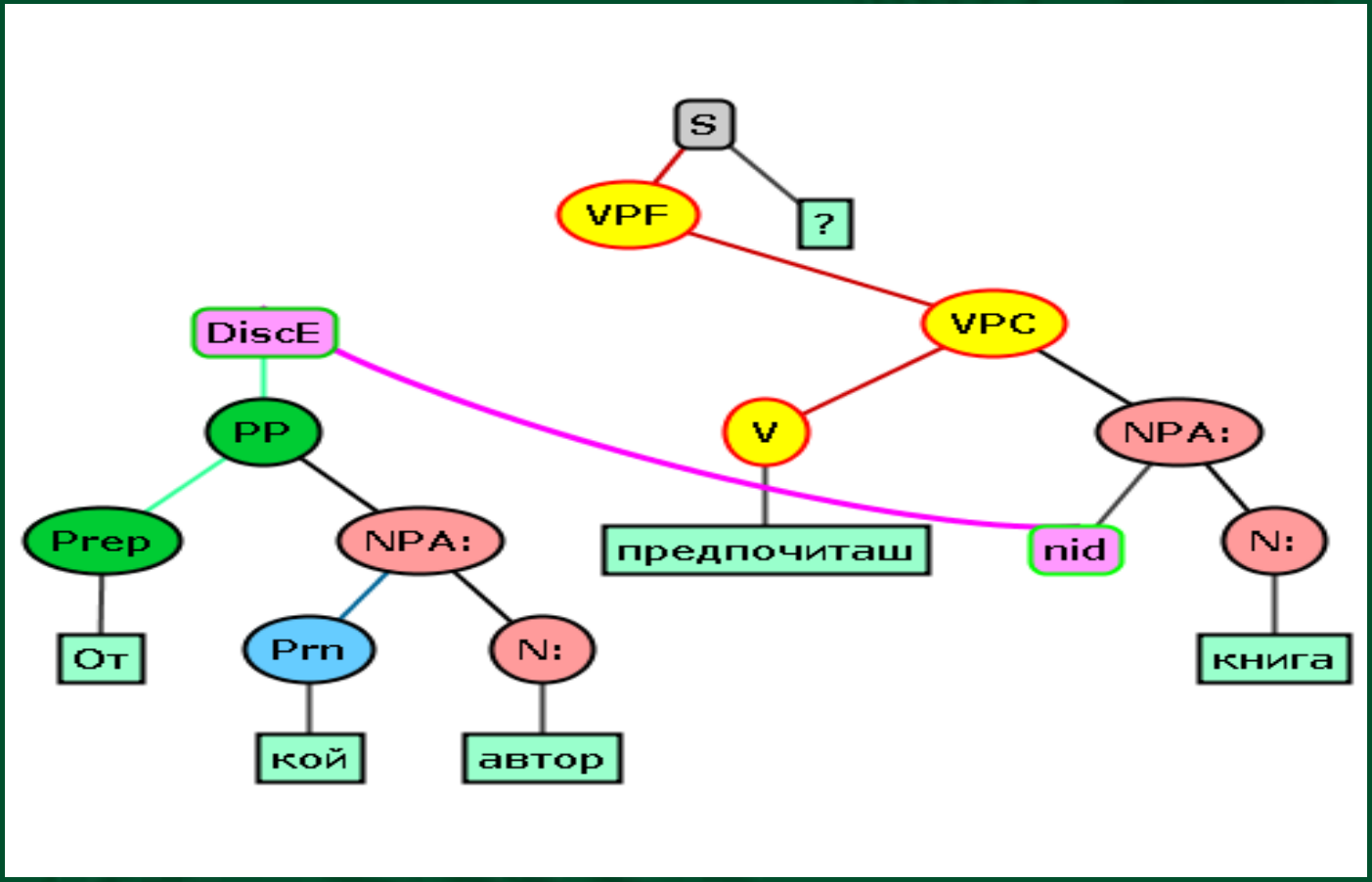
Прага посещава Петя

Петя Прага посещава

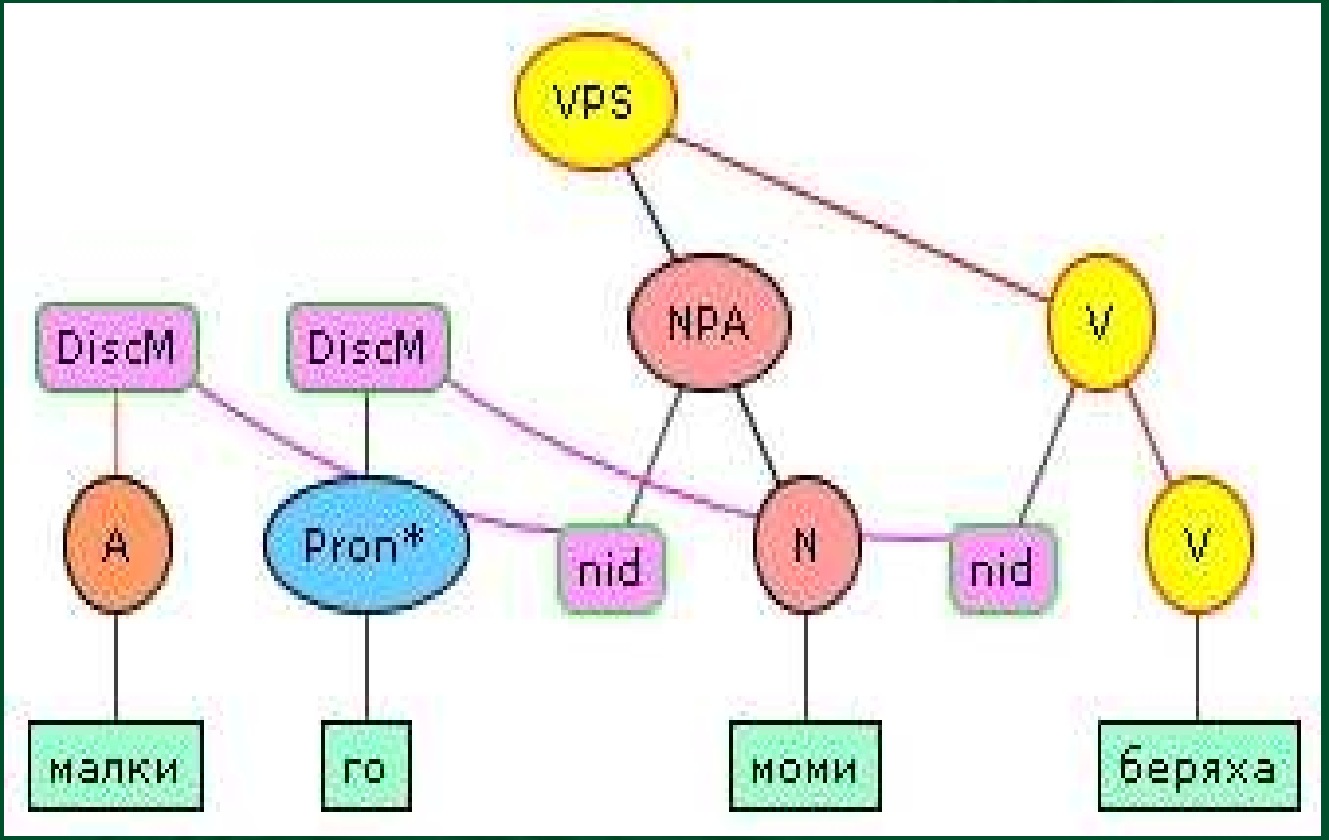
Discontinuities: higher dependant realized between the head and a lower dependant



Discontinuities: outer realization of an inner element (the head is lower) (extraction)



Discontinuities: the elements of two constituent structures are mixed (rare)

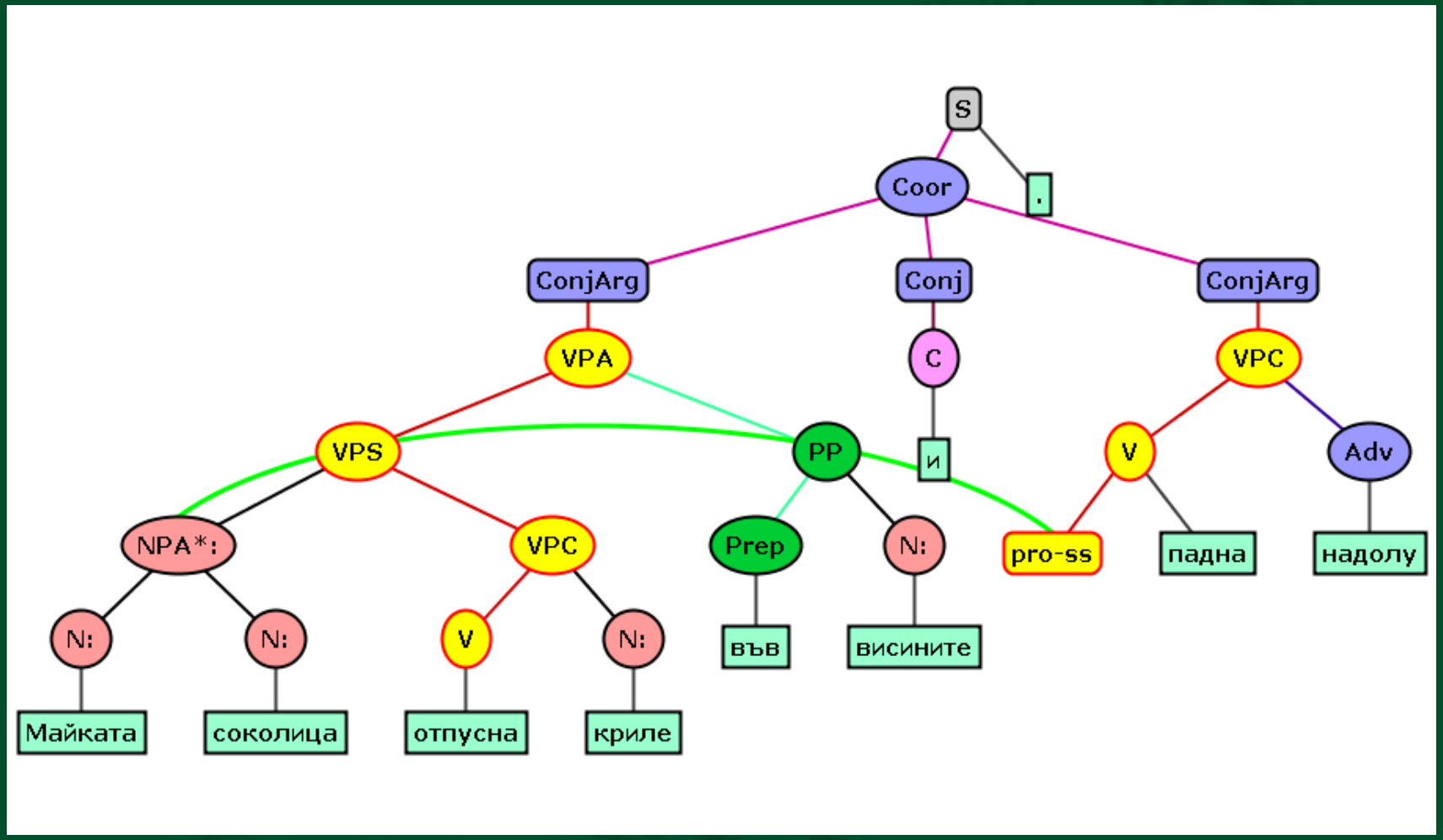


Coordination

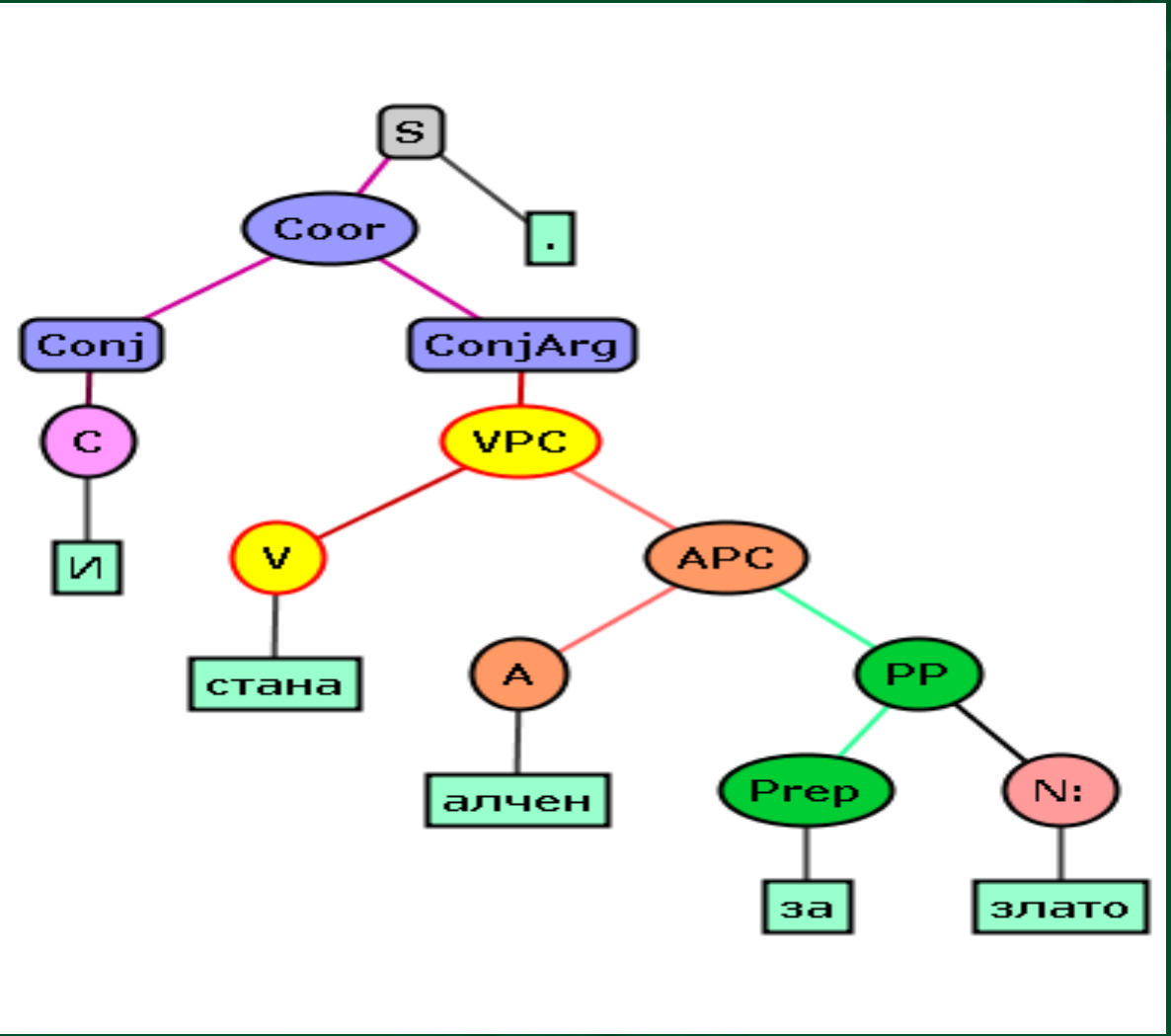
The decisive factors:

- the grammatical role of the conjuncts, not their syntactic label
- the valence of the conjuncts
- the order of dependants realization

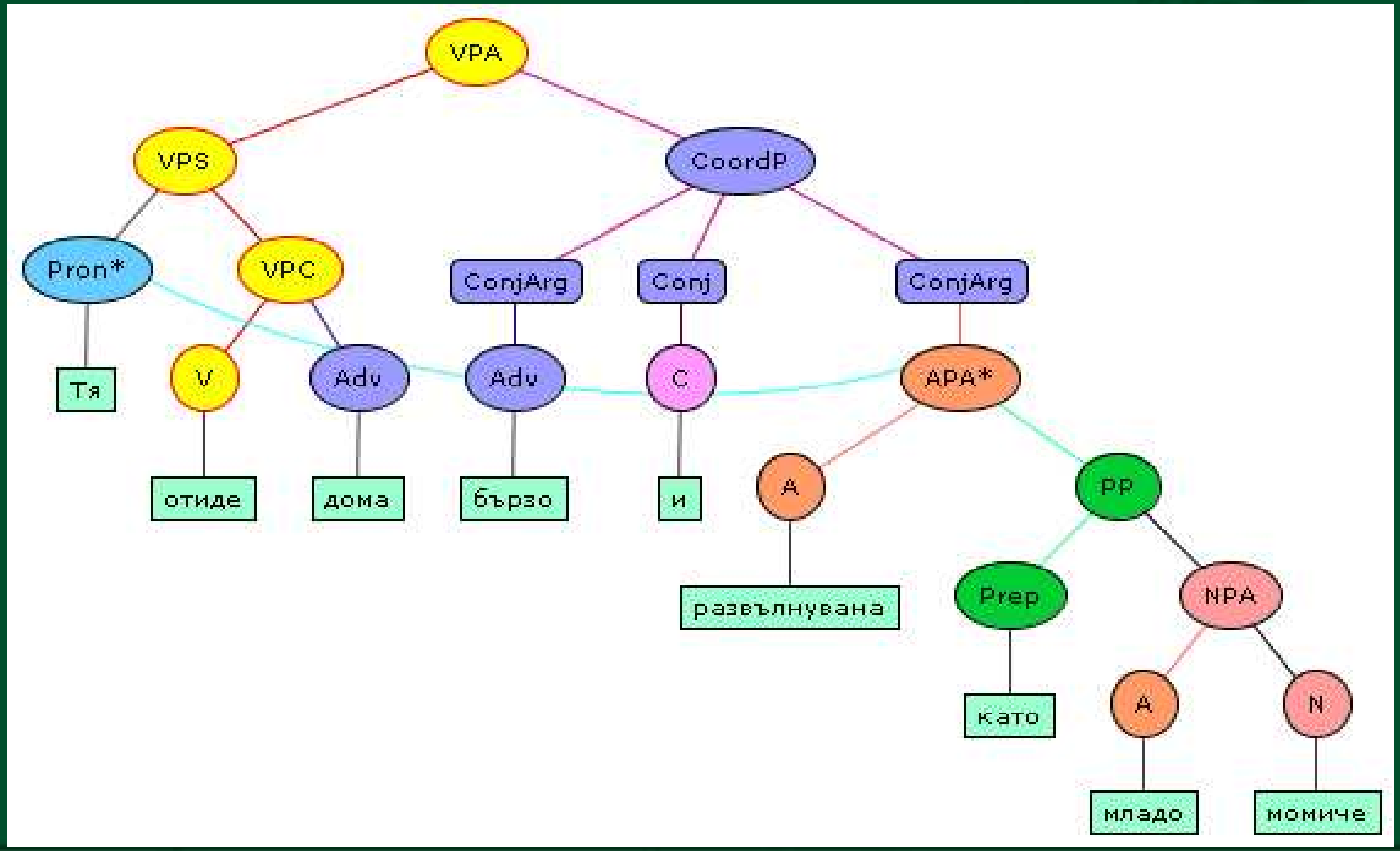
Coordination: the realization of dependants matters



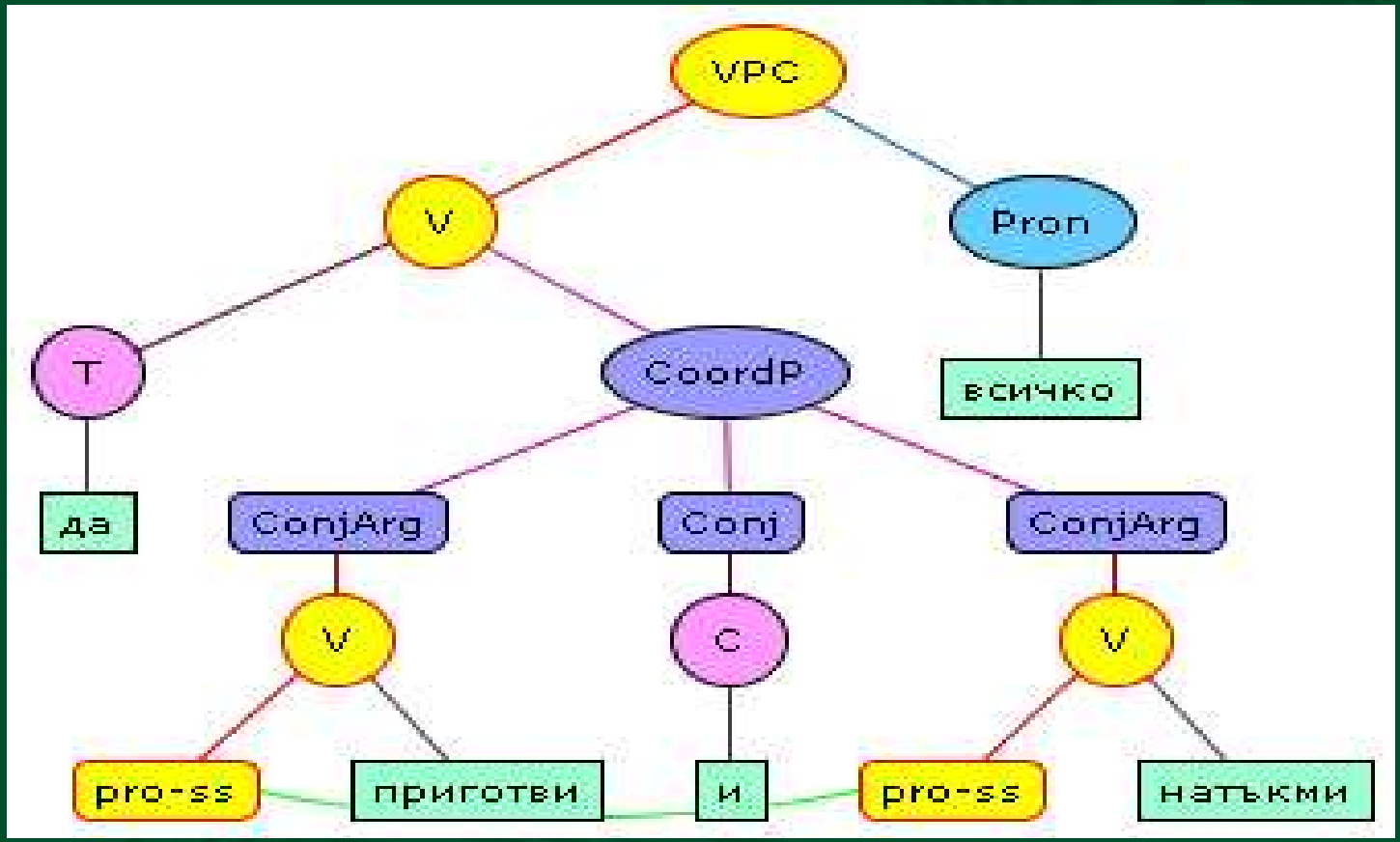
Coordination: one conjunct



Coordination: grammatical role matters



Coordination: valence of the conjuncts matters



Ellipsis

- Repaired within the sentence

  V-Elip, N-Elip, Prep-Elip, PP-Elip

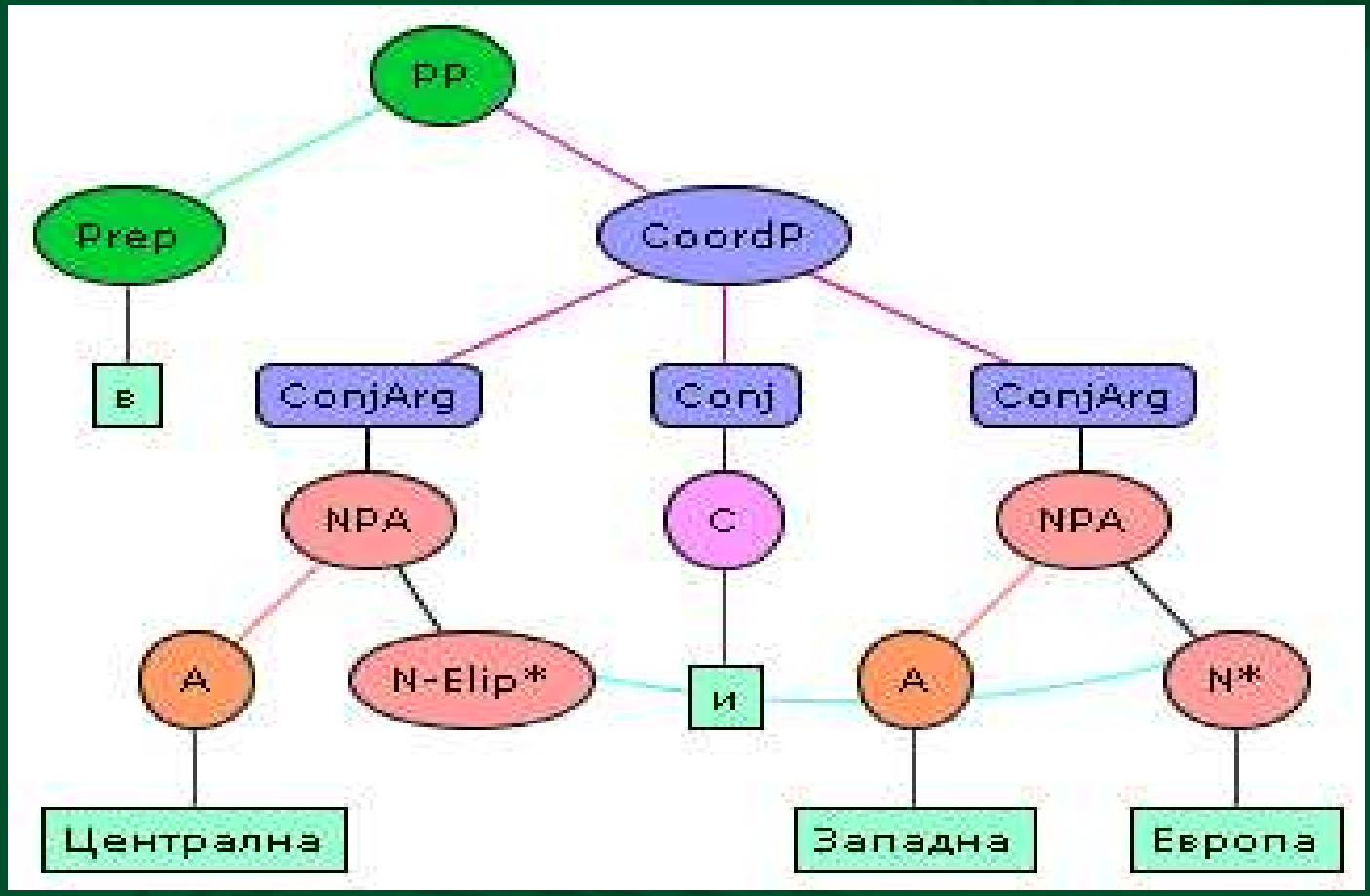
  Attributes: equal, variant, negation

- Depending on broader context

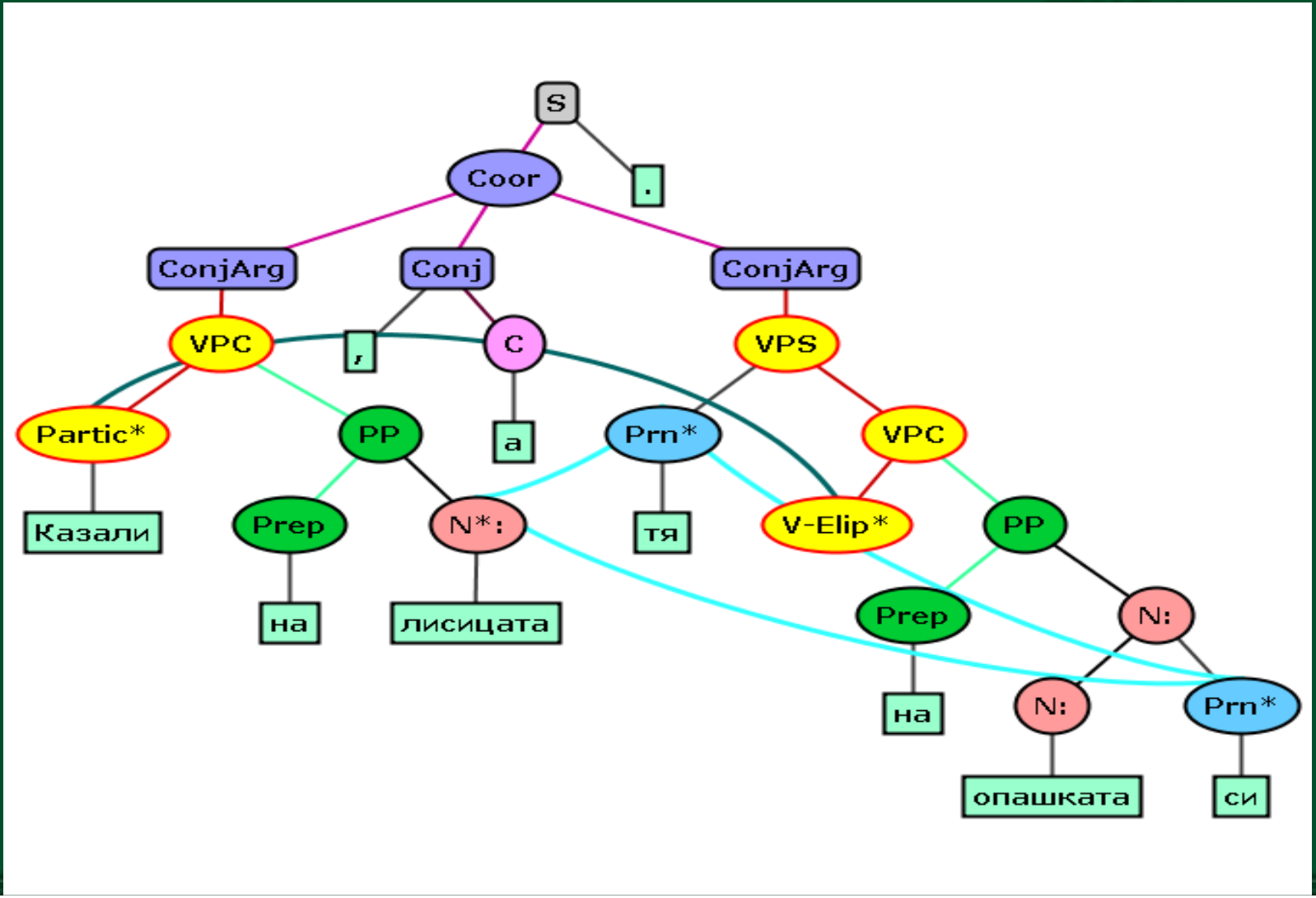
  VD-Elip, ND-Elip, PPD-Elip

  Attributes: world knowledge, discourse, exists (only for VD-Elip)

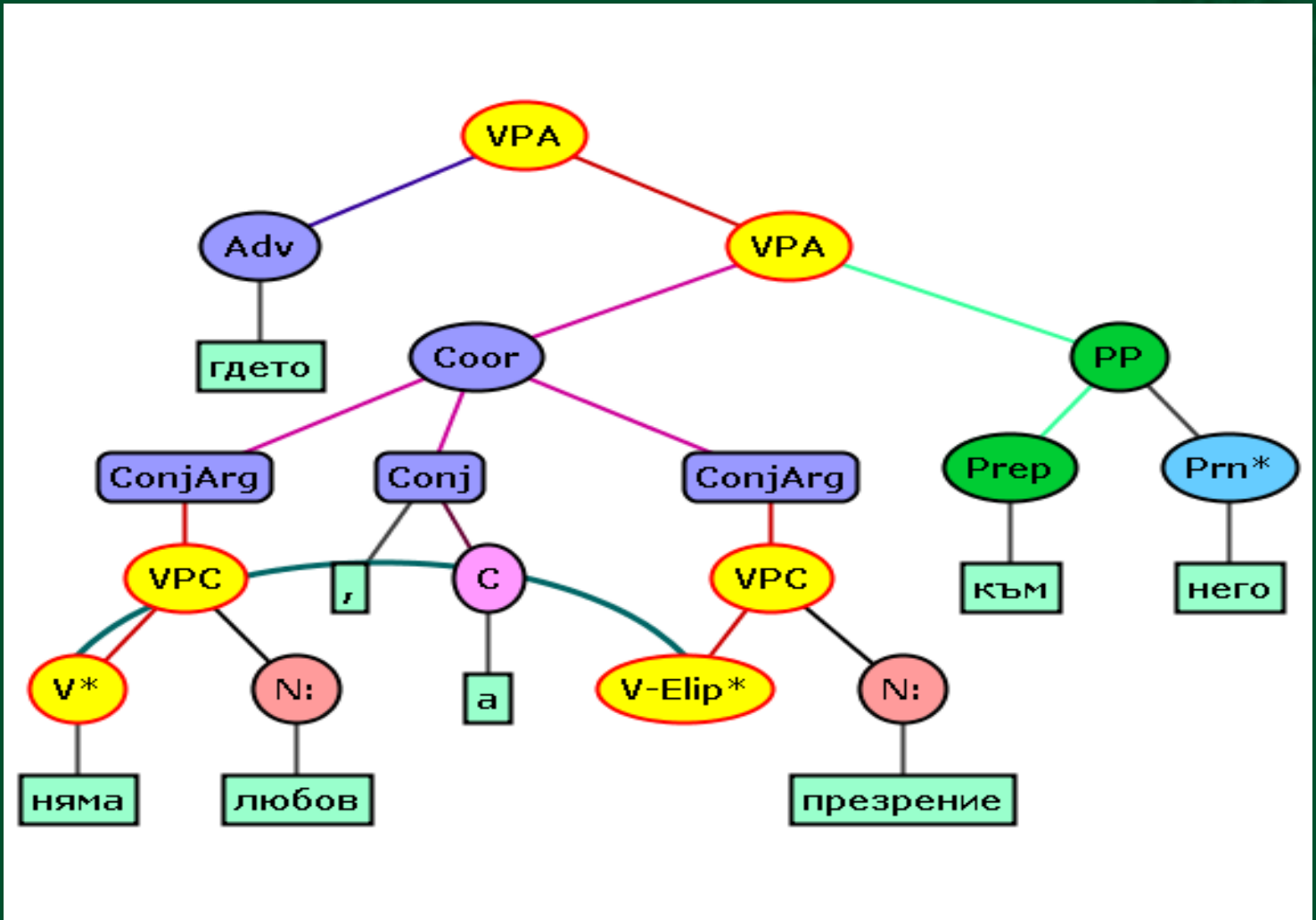
Ellipsis within the sentence: attribute=equal



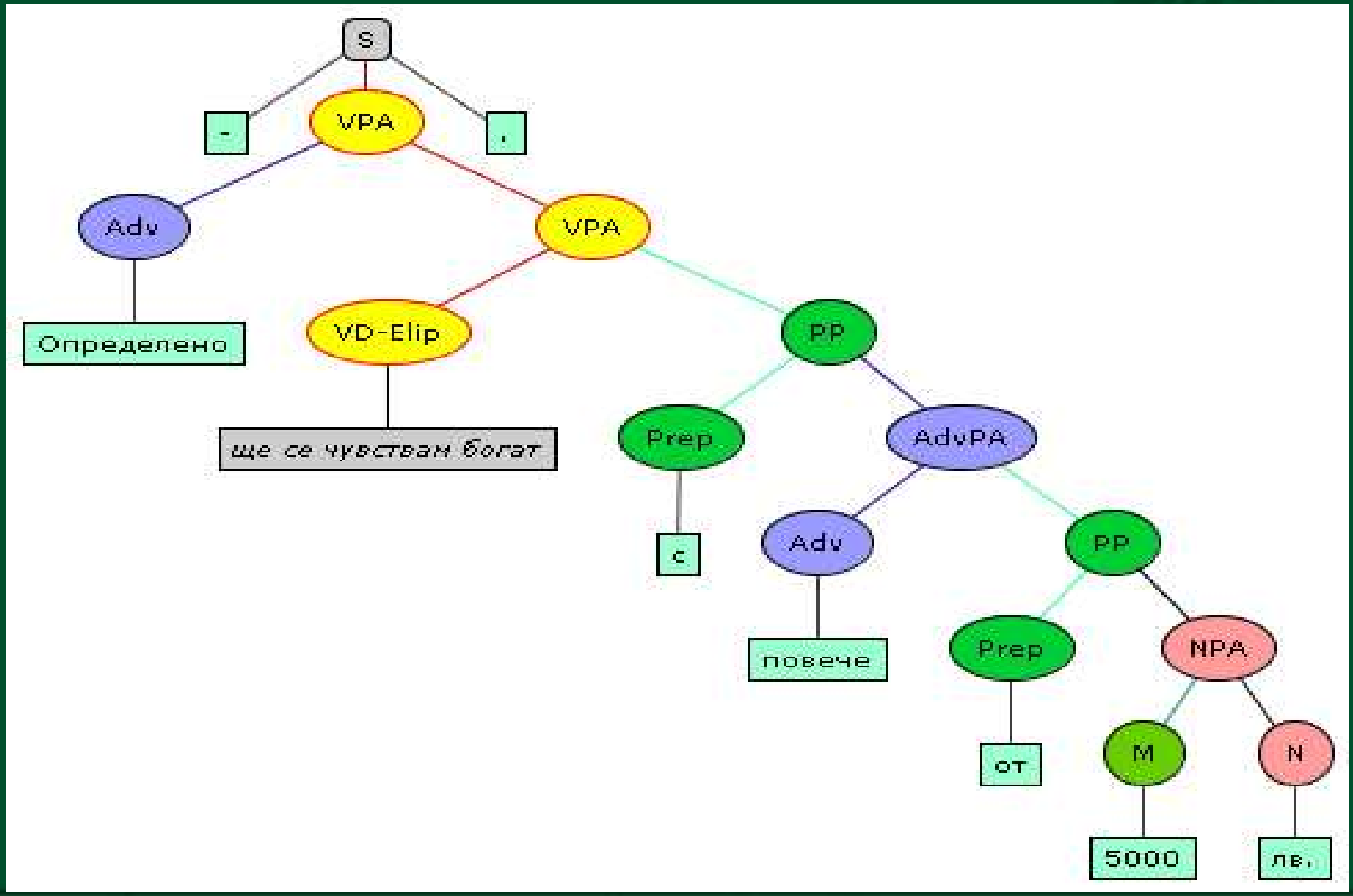
Ellipsis within the sentence:
attribute=variant



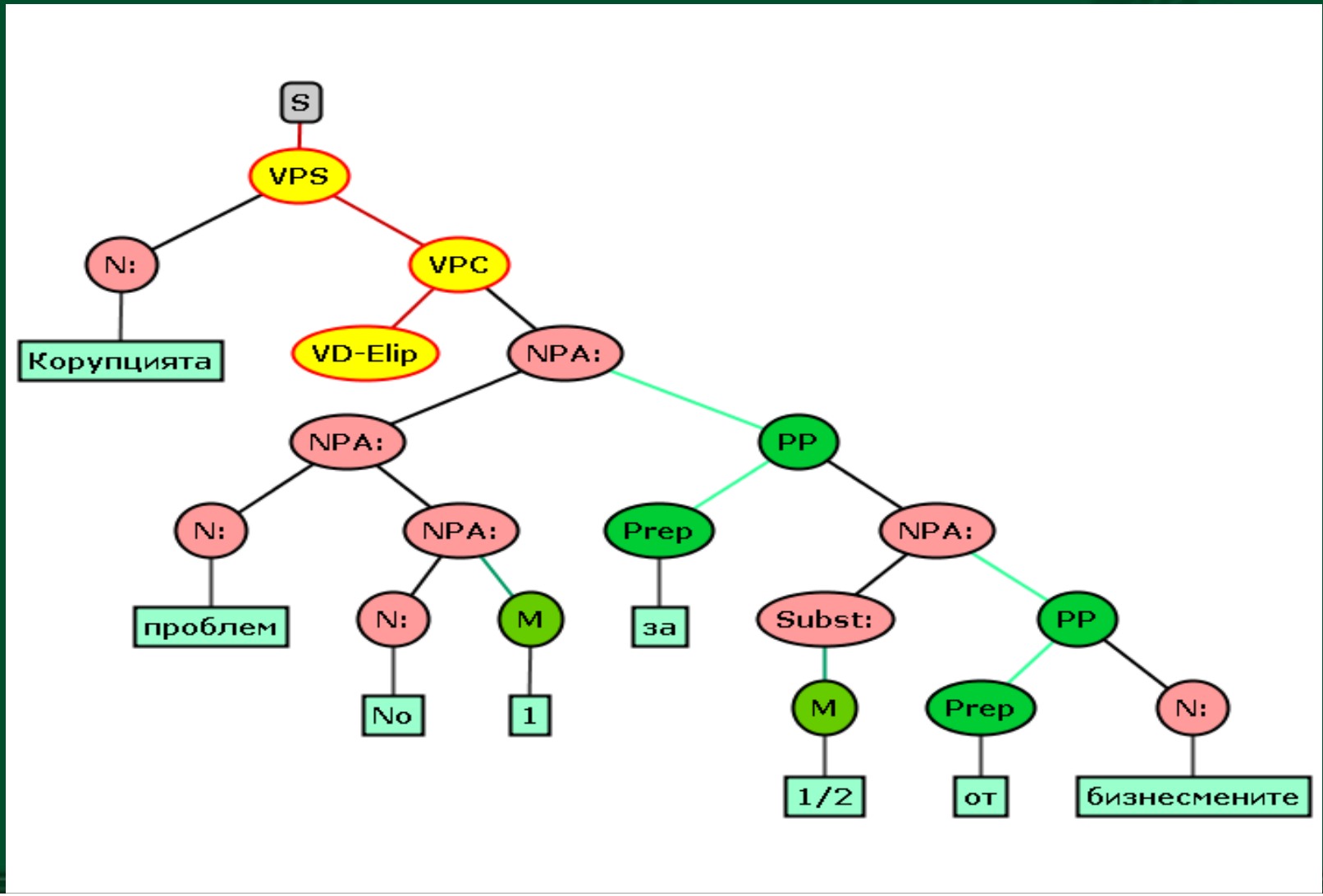
Ellipsis within the sentence:
attribute=negation



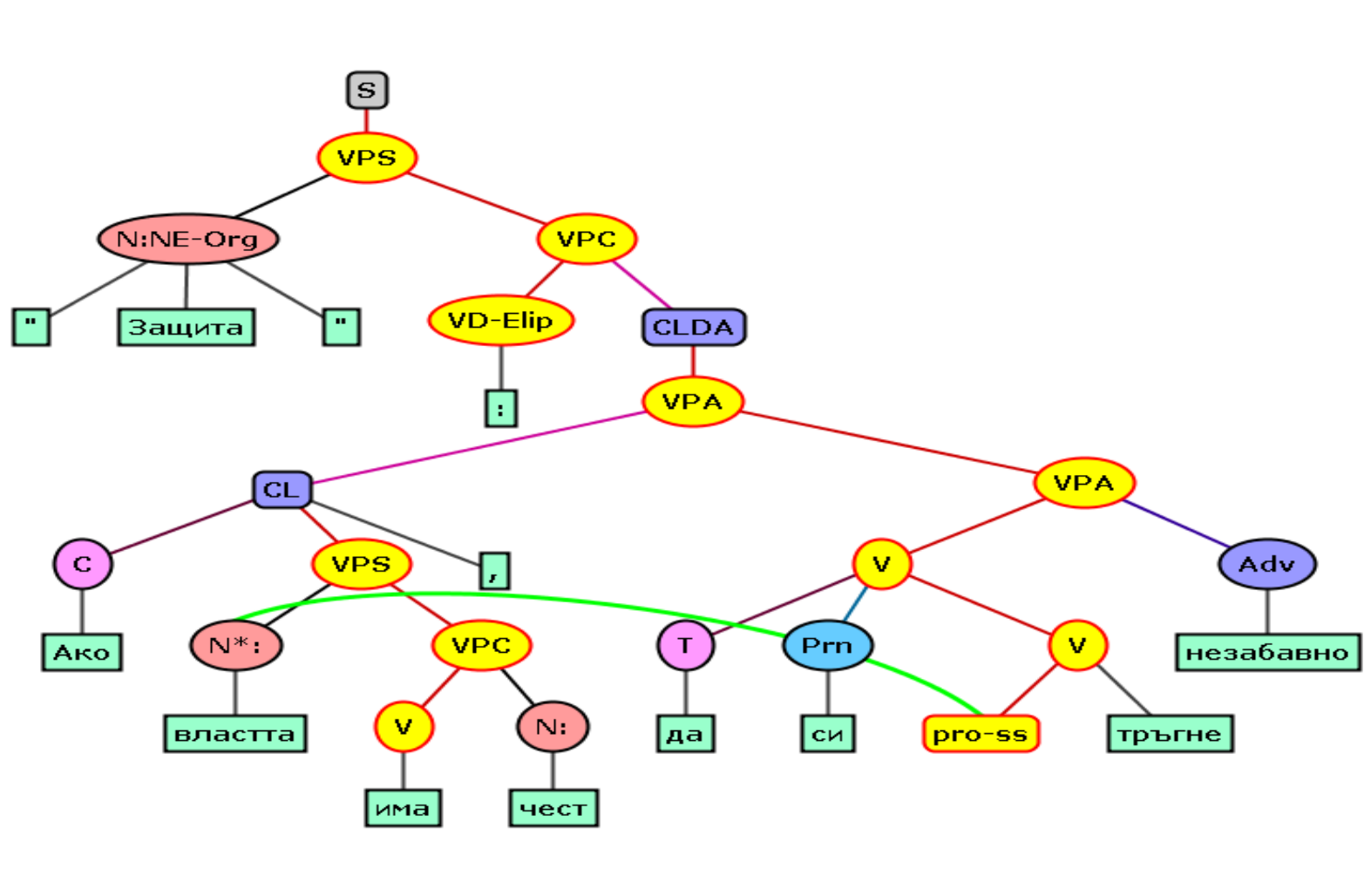
Ellipsis within broader context: attribute=discourse



Ellipsis within broader context: attribute=exists



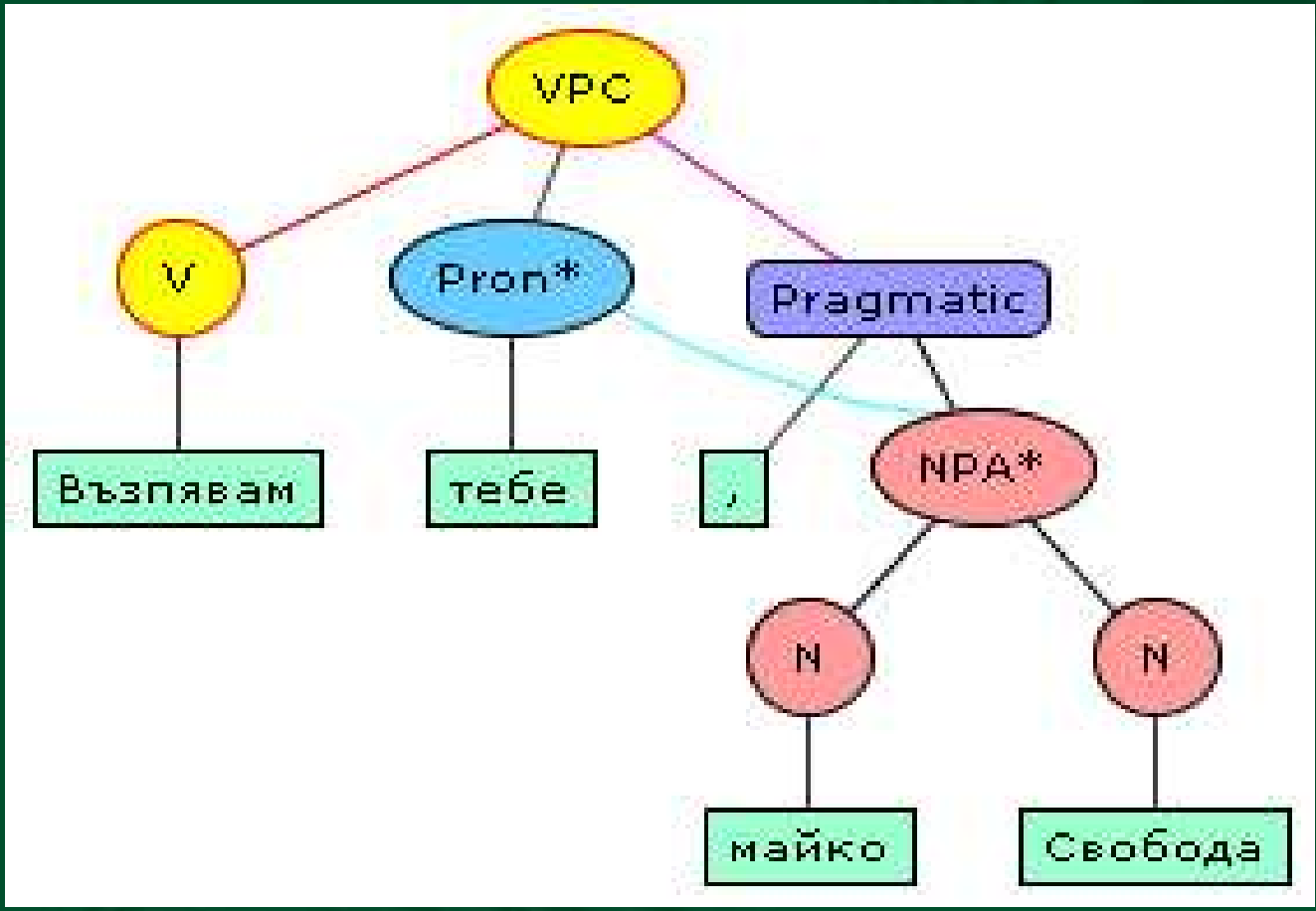
Ellipsis within broader context: attribute=world knowledge



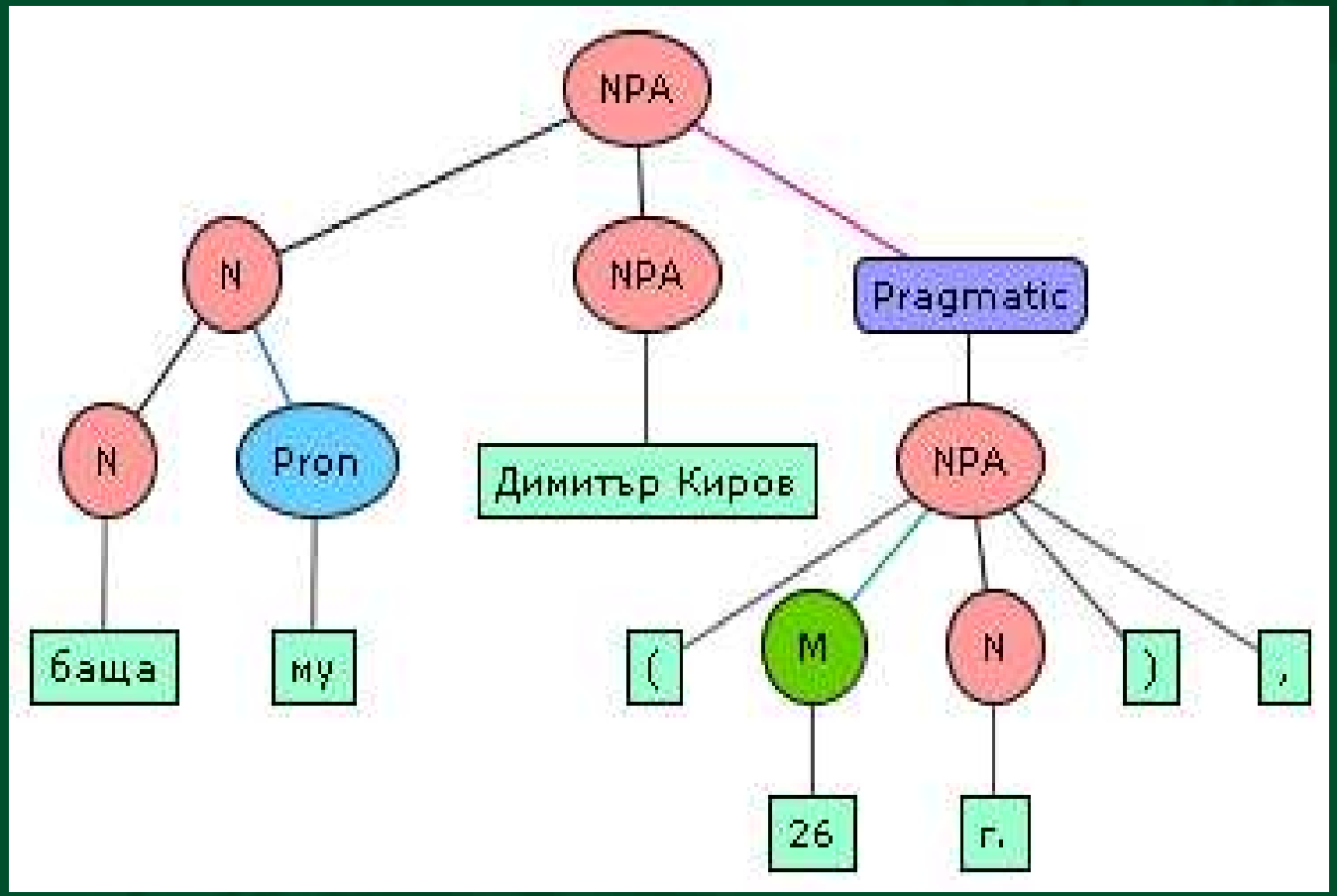
Pragmatic expressions

- Vocatives
- Modal adverbials
- Parenthetical elements
- Foculizers

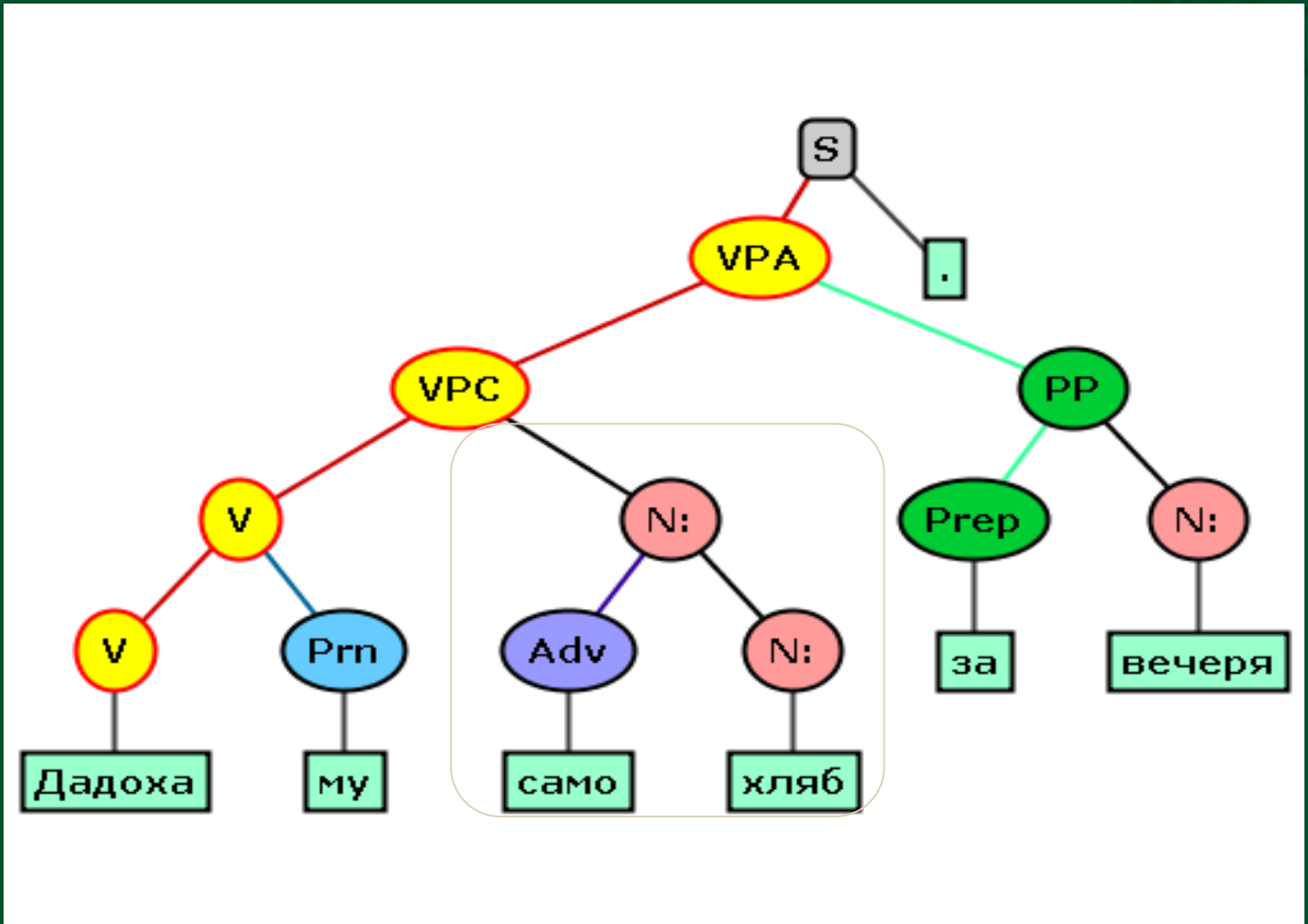
Vocatives



Parenthetical elements



Foculizer



Coreference

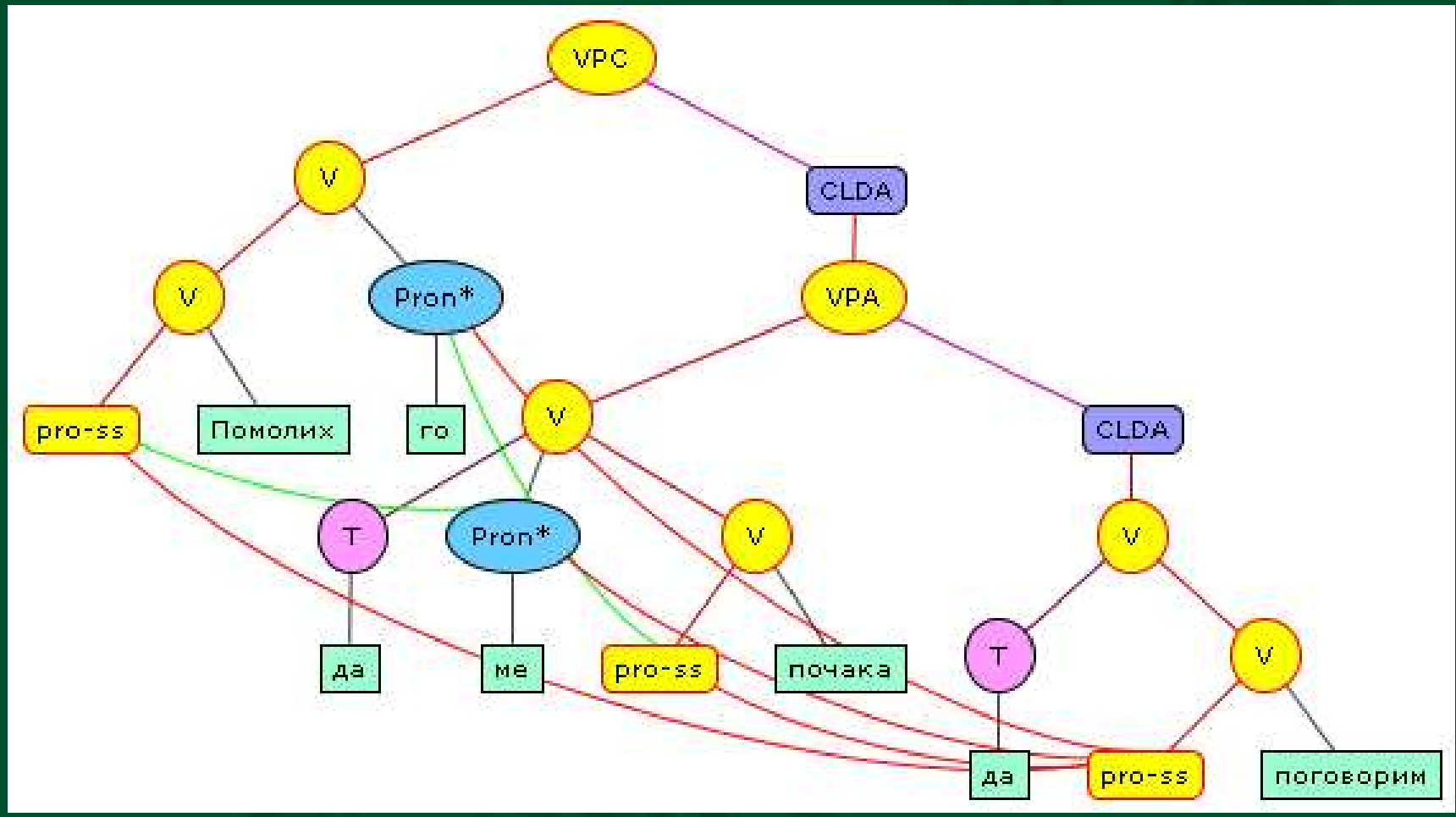
Types:

- equality
- member of a set
- subset of a set

Linguistic parameters:

- pro-dropness
- secondary predication
- binding

Coreference: member of, pro-dropness, control



Next steps in the treebank development

- Towards a Semantic Lexicon connected to an Ontology (SIMPLE project)

Phase 2 consists of two main tasks:

- shallow semantic annotation, i.e. explication of semantic parameters of the head and its dependants
- extension of co-reference relations, i.e. adding more types of relations to the existing ones (part of, bridging ones)

Conclusions

- Bulgarian Treebank exists in two formats now: HPSG-based and Dependancy-based
- It will be expanded in two directions:
  - Quantity (more sentences)
  - Semantic annotation

Accessability

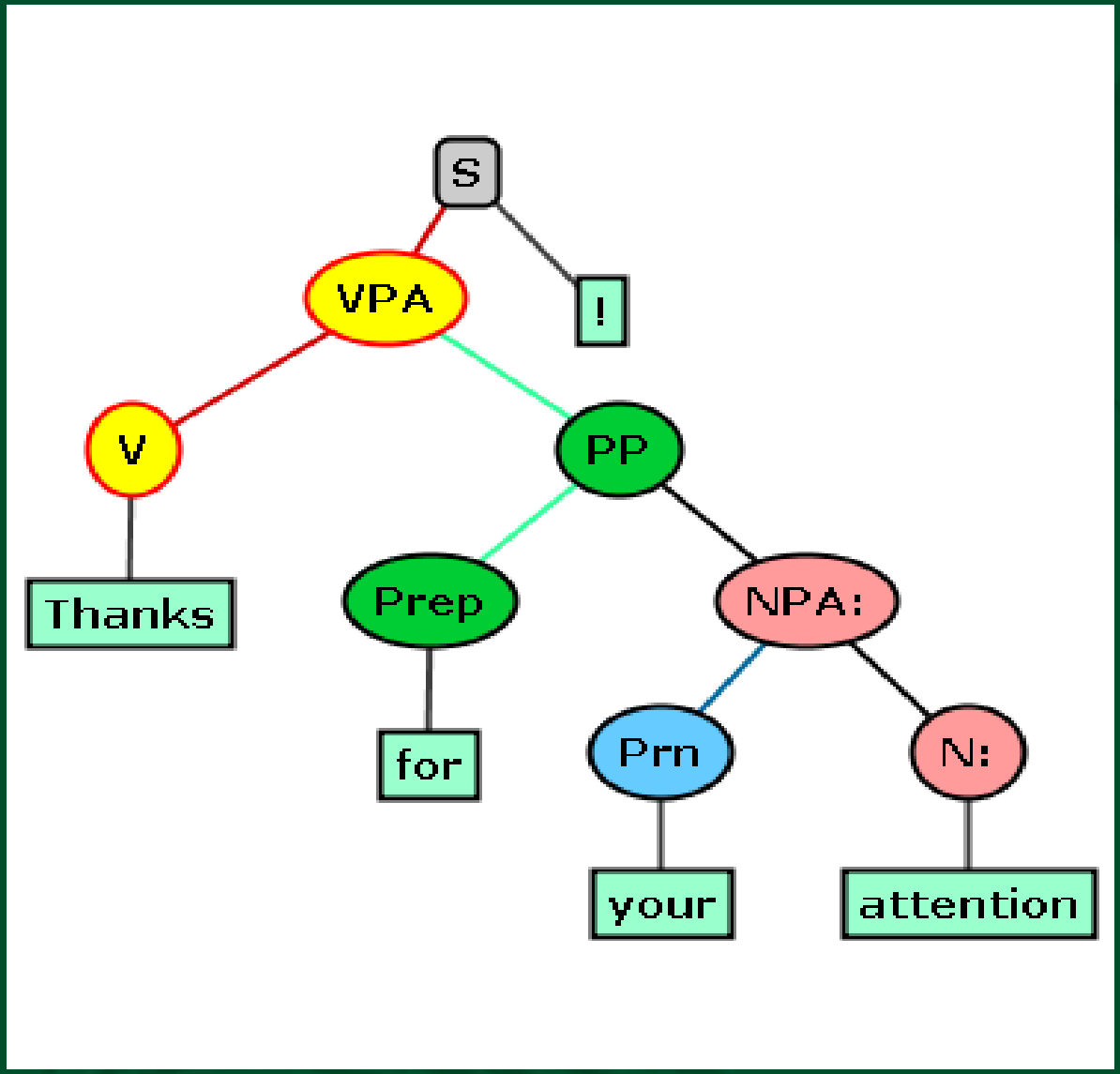
Link: www.bultreebank.org

Free use of:

- CLaRK system
- Dependancy format of Bulgarian Treebank
- Morphologically processed corpus of Bulgarian

Facilities:

- Technical Reports
- Documentation


S
VPA
!
V
PP
Thanks
Prep
NPA:
for
Prn
N:
your
attention