Bulgarian Language Resources and Technology for Deep Grammar Machine Translation

Kiril Simov
Linguistic Modelling Laboratory
Institute for Parallel Processing
Bulgarian Academy of Sciences

MOLTO Project Meeting
8th September, 2010, Varna, Bulgaria

Plan of the Talk

- Introductory Notes
- Bulgarian HPSG-based Treebank
- Bulgarian Ontology-based Lexicon
- Bulgarian Language Technology
- HPSG-based Statistical Translation
- Conclusion

Institute for Parallel Processing

- **BulTreeBank** – An HPSG-based treebank of Bulgarian.
- **BulTreeBank Text Archive** – Texts annotated up to paragraph level with respect to TEI guidelines (near 400 million words)
- **BulTreeBank Morphosyntactic Corpus** – Annotated with grammatical information
- **Bulgarian CLEF Corpus** – Supporting the evaluation of question answering and information retrieval systems
- **Bulgarian LT4eL Corpus** – Grammatical/Semantic annotation.
- **Morphological Dictionary of Bulgarian**.
- **BulTreeBank Gazetteers** – Lexicon of proper names
- **BulTreeBank Partial Grammar** – simple NPs and VPs
- **Dependency Parser for Bulgarian**.

Institute of Mathematics and Informatics

Corpora

- MULTEXT-East Multilingual Parallel Annotated and Aligned Corpus
- MULTEXT-East Comparable Corpora: BG fictions, BG news
- Bulgarian-Polish Parallel Annotated Corpus
- Bulgarian-Polish Comparable Corpus
- Bulgarian-Polish-Lithuanian Parallel and Comparable Corpora

Language-specific Resources

- MULTEXT-East Language-specific Resources - TEI-compliant Morphosyntactic Specifications for Corpora and Lexicon encoding
- Bulgarian Lexicon
- Bulgarian Corpus
- Bulgarian LDB for integrated multilingual CONCEDE LDBs

Bilingual digital dictionaries

- Bulgarian-Polish online dictionary
- LDBs for Bulgarian-Lithuanian online dictionary (in progress)
- Slovak-Bulgarian Terminology DB (in progress)

Institute for Bulgarian Language

- **Bulgarian WordNet**
- **Grammar Dictionary of Bulgarian** - An Electronic Grammar Dictionary of Bulgarian
- **Automatic spelling checking system: ItaEst**
- **Bulgarian written corpus**
- **Tagged corpus of Bulgarian** - The Tagged Corpus is the result of the manual POS disambiguation of each wordform
- **Semantic Corpus of Bulgarian** - The Semantic Corpus contains sense-disambiguated lexical items defined in the context of occurrence

Plovdiv University “Paisii Hilendarski”

- **Bulgarian WordNet** (with IBL)
- **Dictionary of Bulgarian Inflection Morphology**
- **Bulgarian POS Tagger**
- **Chunker of Bulgarian**

BulTreeBank

Based on Kiril Simov, Petya Osenova, Alexander Simov, Milen Kouylekov. *Design and Implementation of the Bulgarian HPSG-based Treebank*. Special Issue on Treebanks and Linguistic Theories. Research on Language & Computation. Springer Science+Business Media B.V. Volume 2, Number 4.

Work done in projects: BulTreeBank

Goals

- A set of Bulgarian sentences marked-up with detailed syntactic information
- A core set of sentences will be designated inside the treebank
- Reliable partial grammar for automatic parsing of phrases in Bulgarian
- Software modules for compiling, manipulating and exploring the treebank

Requirements for the Annotation

- Adequate representation of the linguistic facts
  - Theory dependency
- Adequate representation of partial and complete analysis
  - Easy transfer of the information
- Convenience for manual annotation
  - minimal information input

Why Theory Dependency? (1)

- On a certain level of granularity the annotation scheme becomes very complicated to be processed consistently
- On a certain level of granularity some linguistic theory has to be exploited
- Two choices:
  - A new “annotation” linguistic theory to be developed, or
  - A well-established existing theory to be adopted

We have chosen HPSG as a base for our treebank

Why Theory Dependency? (2)

- HPSG is one of the major linguistic theories based on rigorous formal grounds
- HPSG allows for a consistent description of linguistic facts on every linguistic level: syntactic, semantic and others
- HPSG allows for different levels of generalisation and therefore enables different experts to work on different levels of analysis
- The formal basis of HPSG allows translation to other formalisms
- There are universal HPSG principles that can be used to support the work of the annotators

Core Set of Sentences

- In the process of the treebank compilation it plays double role
  - Gold standard: this set has to cover the basic linguistic phenomena in Bulgarian
  - HPSG Grammar development basis
- Here we present and discuss the annotation scheme for the treebank

HPSG Language Model

- Linguistic objects
  - Represented as directed graphs (feature structures)
- 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 Linguistic Objects

- The main linguistic object is of sort sign with three main attributes: PHON, SYNSEM and DTRS (for phrases)
- The co-reference is the basic mechanism for ensuring the correct object structure
- The attribute DTRS determines the variety of constituent structures and the grammatical functions

The Hierarchy of Phrases

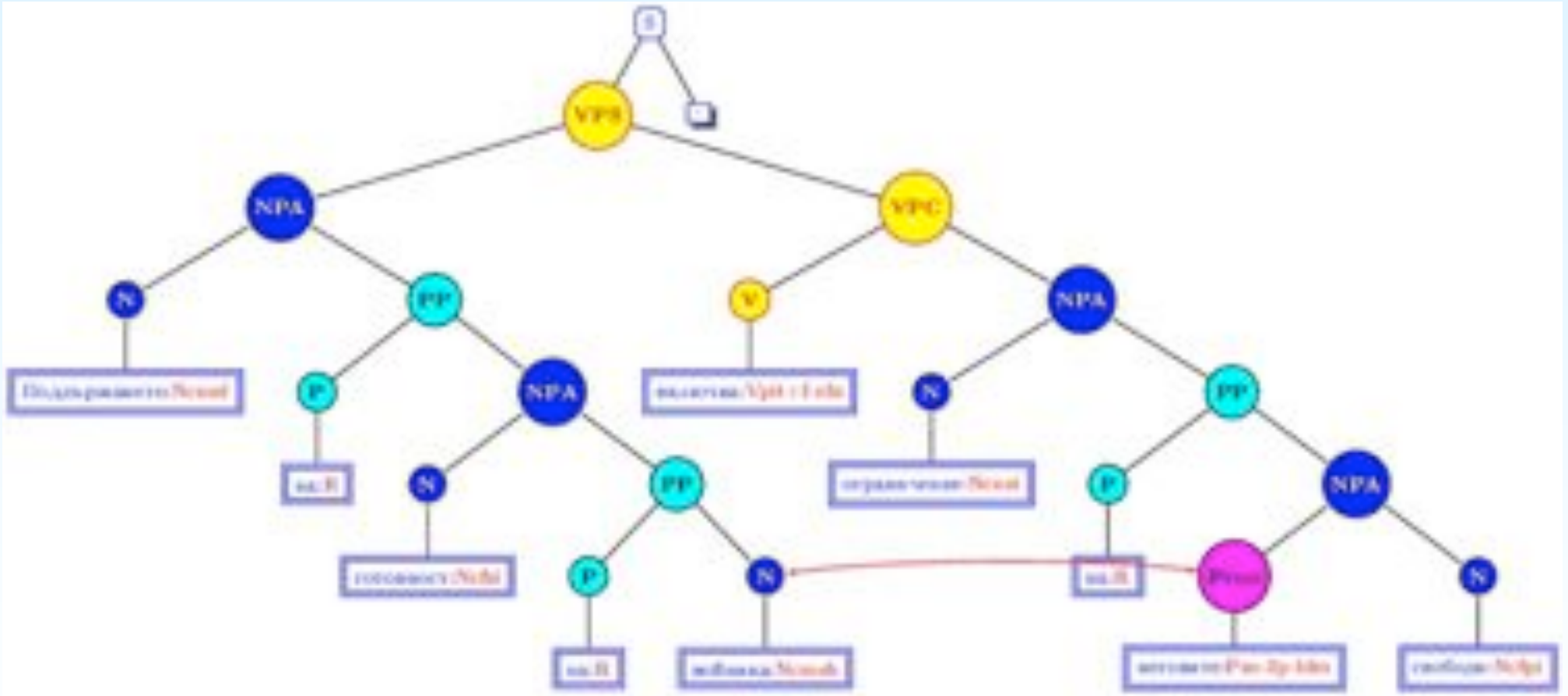
- headed-phrase
  - head-complement
  - head-subject
  - head-adjunct
    - head-sem-adjunct
    - head-pragmatic-adjunct
  - head-filler
- non-headed-phrase

Constituency and Dependency (1)

- 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

Constituency and Dependency (2)



Linguistic Object Representation

The representation of the linguistic objects (of sort *sign*) in the core set of sentences is based on:

- Context-free-like trees
- Coreferencial relations over the trees
- Node labels reflecting the *synsem* information

Word Order and Discontinuity

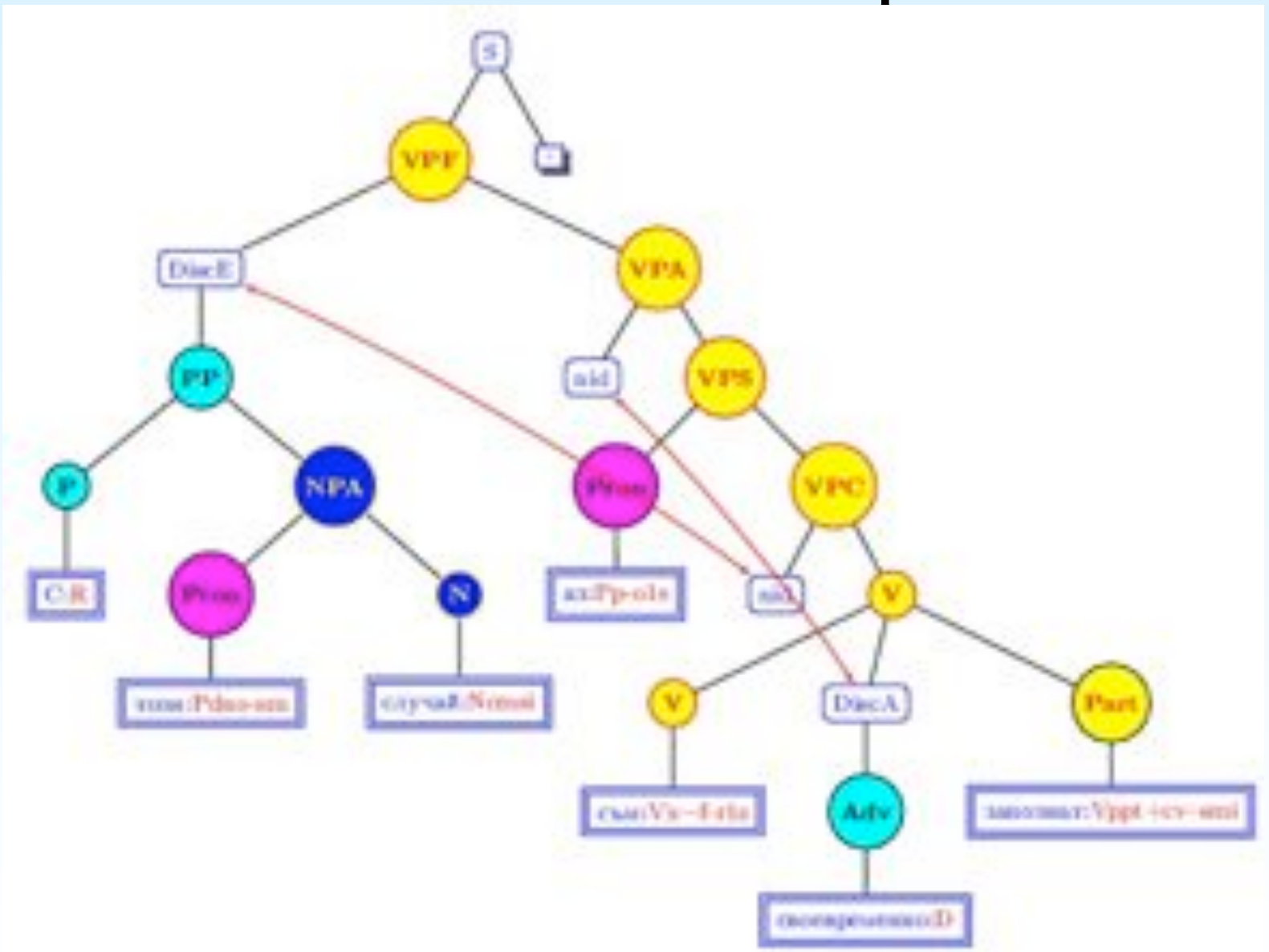
- Continuous realisation of daughters
- Head dependants permutation

  a constituent from an upper level of the hierarchy is realised between constituents of a lower level
- Mixture of two saturated constituents

  the constituents of two saturated phrases are mixed with each other
- External realisation of an inner constituent

  extraction

Realisation of the Dependants



Linguistic Parameters

- We rely on two basic assumptions:
  - We use a domain-phenomena cross-classification, where the main syntactic domains are defined and the phenomena are analyzed
  - We analyze the data according to the following HPSG-oriented criteria: the type of the sign, headedness, the typology of words and phrases, the saturation condition

Core Domains: NP

- Bare Bulgarian NPs are always functionally complete (lexical category N)
- NP dependency structures: head-complement (NPC), head-adjunct (NPA)
- Classification criteria: ontological features (mainly for the named entities), ellipsis, substantivization, nominalization

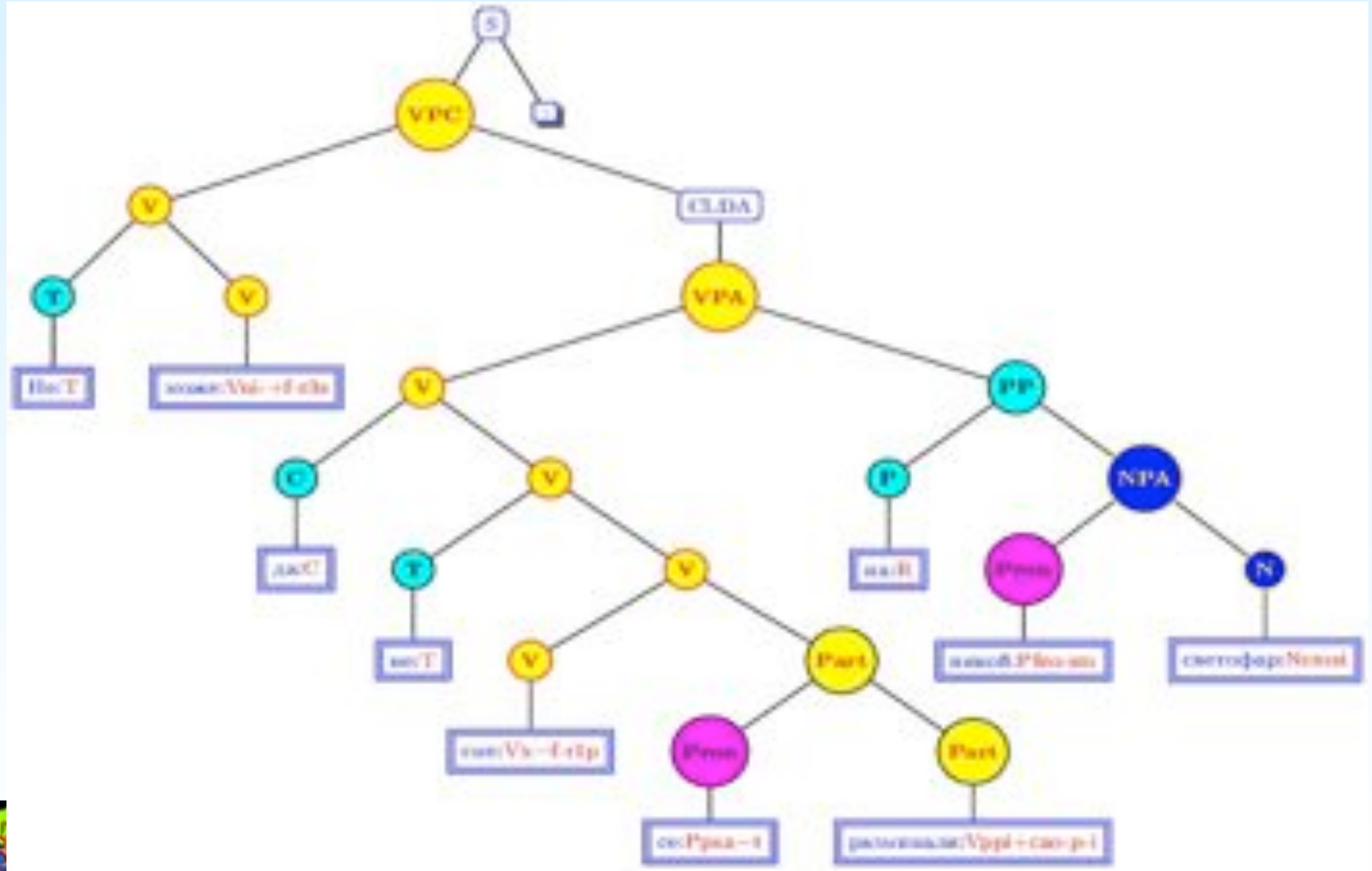
Core Domains: VP (1)

- VPs are classified as lexical and phrasal
- The lexical (V) includes:
  - Bare verbs
  - Verbs with clitics
  - Da-constructions
  - Analytical verb forms
  - Elliptical verbs

Core Domains: VP (2)

- The phrasal category is recursive:
  - First, the verb with its full-fledged complement(s) forms a head-complement phrase (VPC)
  - Then, the head-complement VP takes the subject and forms a head-subject phrase (VPS)
  - The adjuncts are attached last and form head-adjunct (VPA) projections
  - Each extracted element without a structural parent is attached to a head-filler phrase (VPF)
  - CL stands for a saturated verb phrase

Da Clause



Core Domains: AP, AdvP, PP

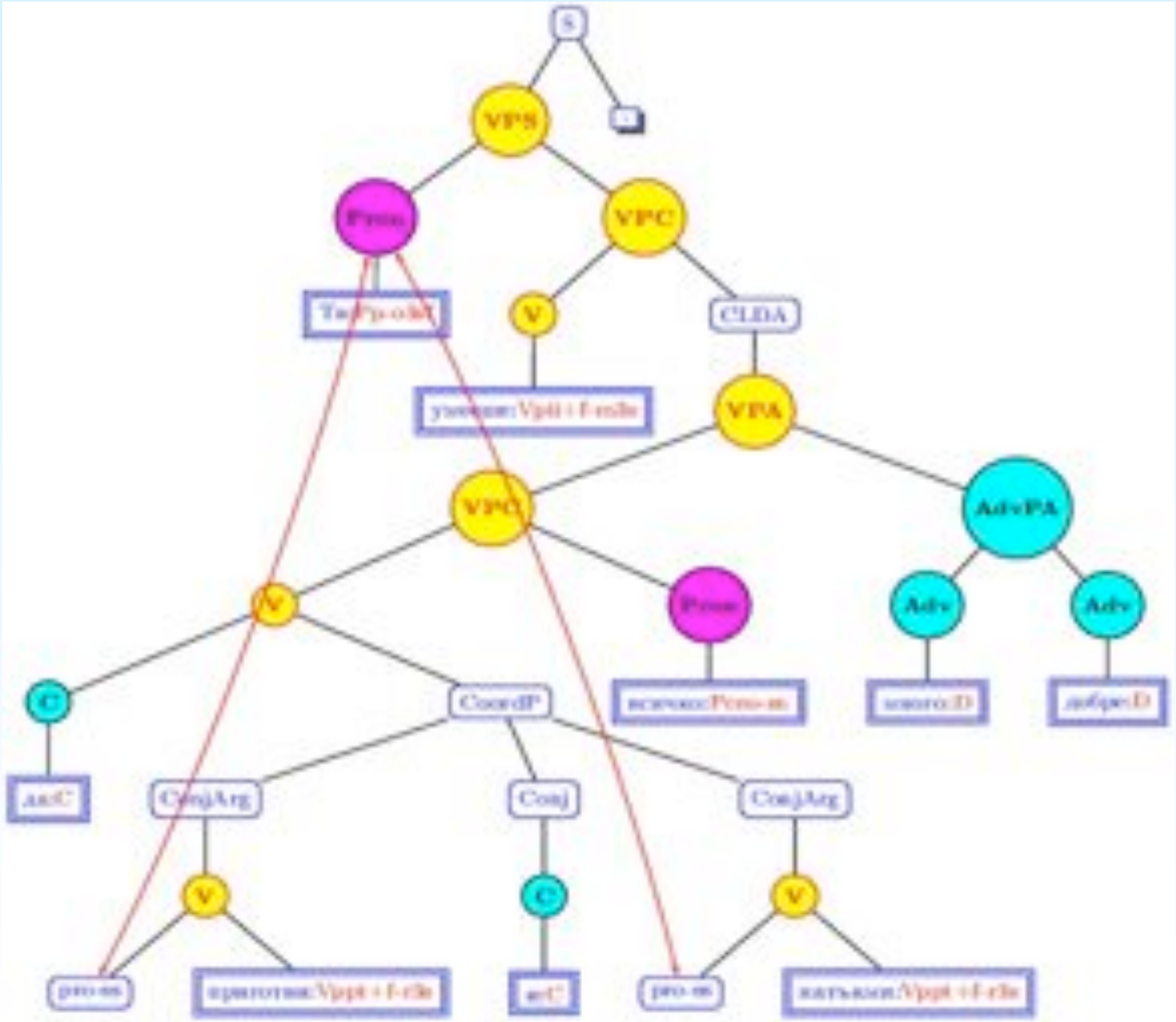
- Lexical adjective (A) can be combined with possessive clitic
- AP can be head-complement phrase (APC), and head-adjunct phrase (APA)
- Non-modified adverb is marked lexically (Adv)
- AdvP can be head-adjunct phrase (AdvPA) and head-complement with a gerund head (AdvPC)
- PP is always a head-complement phrase

Coordination

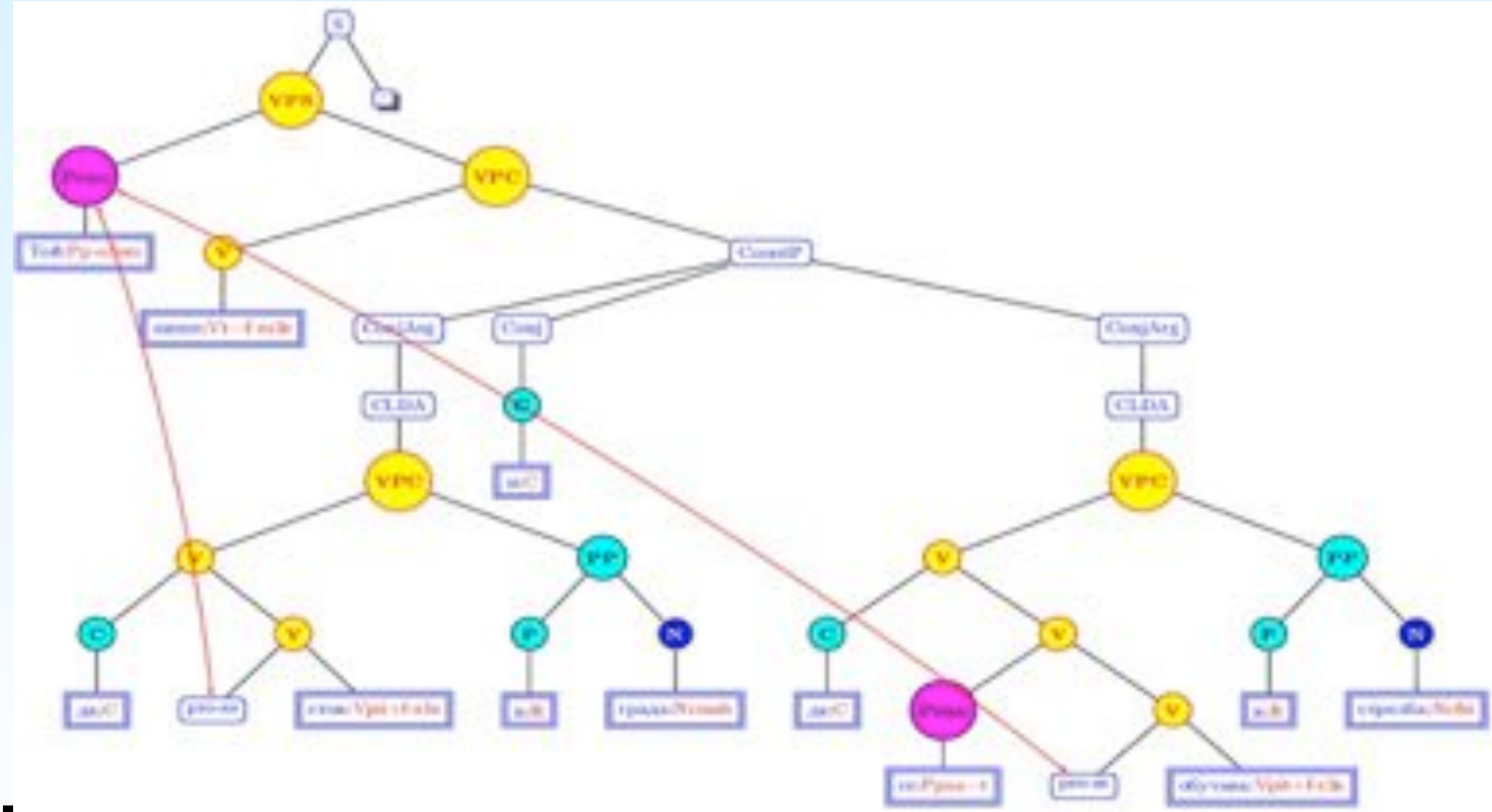
Coordination is treated as a non-headed phrase with the following requirements:

- The conjuncts have to agree in their valency potential: Valency lists and Mod feature
- They can be underspecified with respect to the category: coord

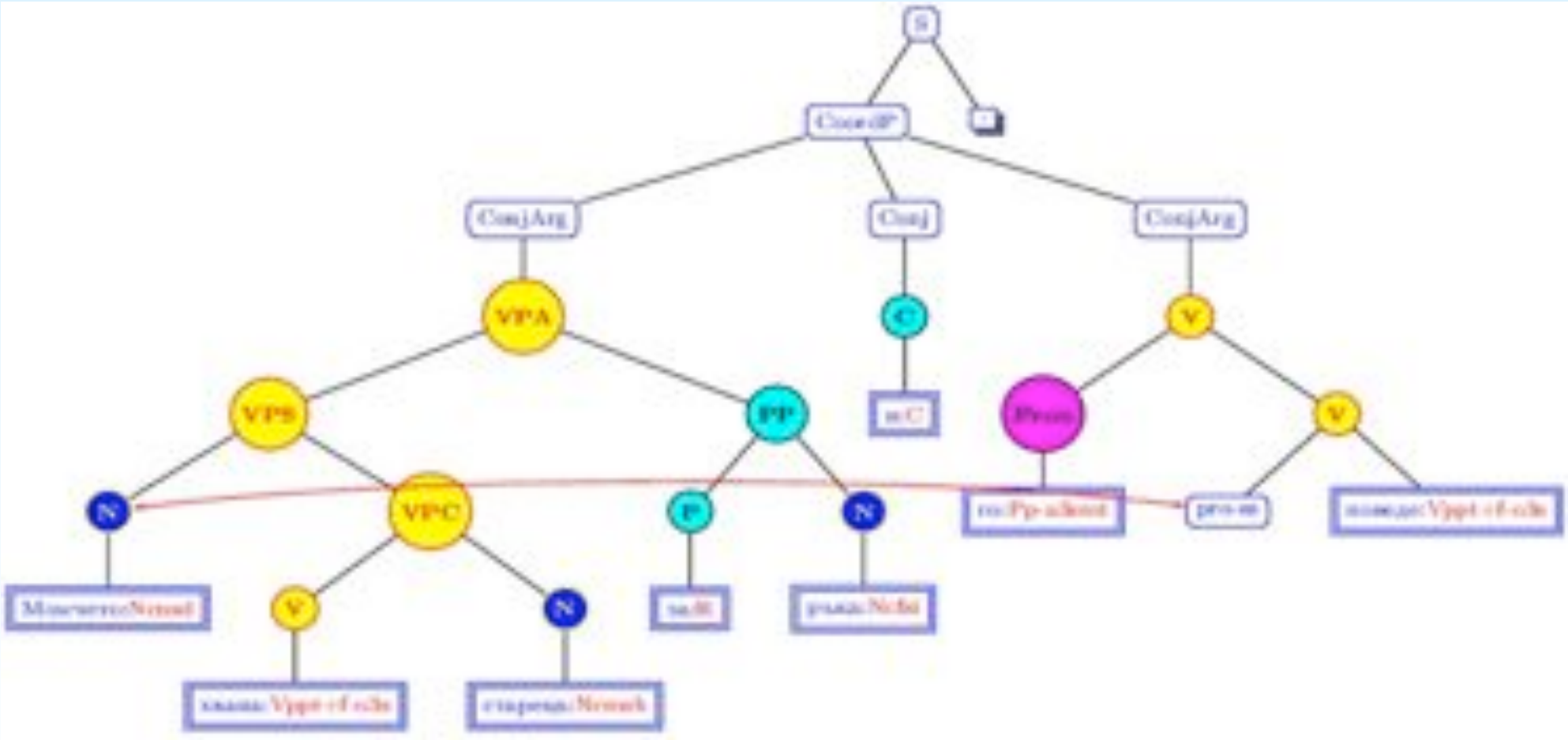
Lexical



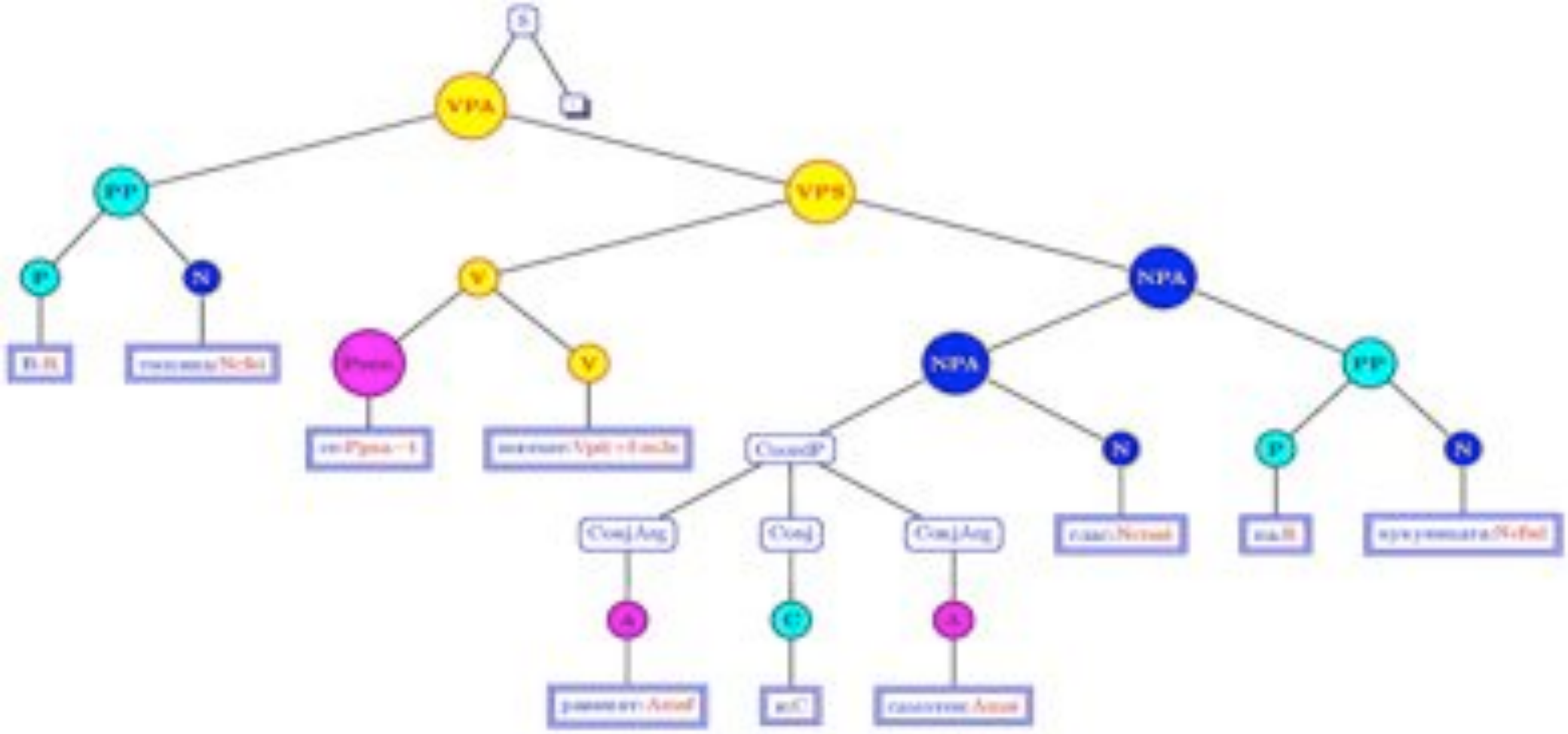
Clausal (1)



Clausal (2)

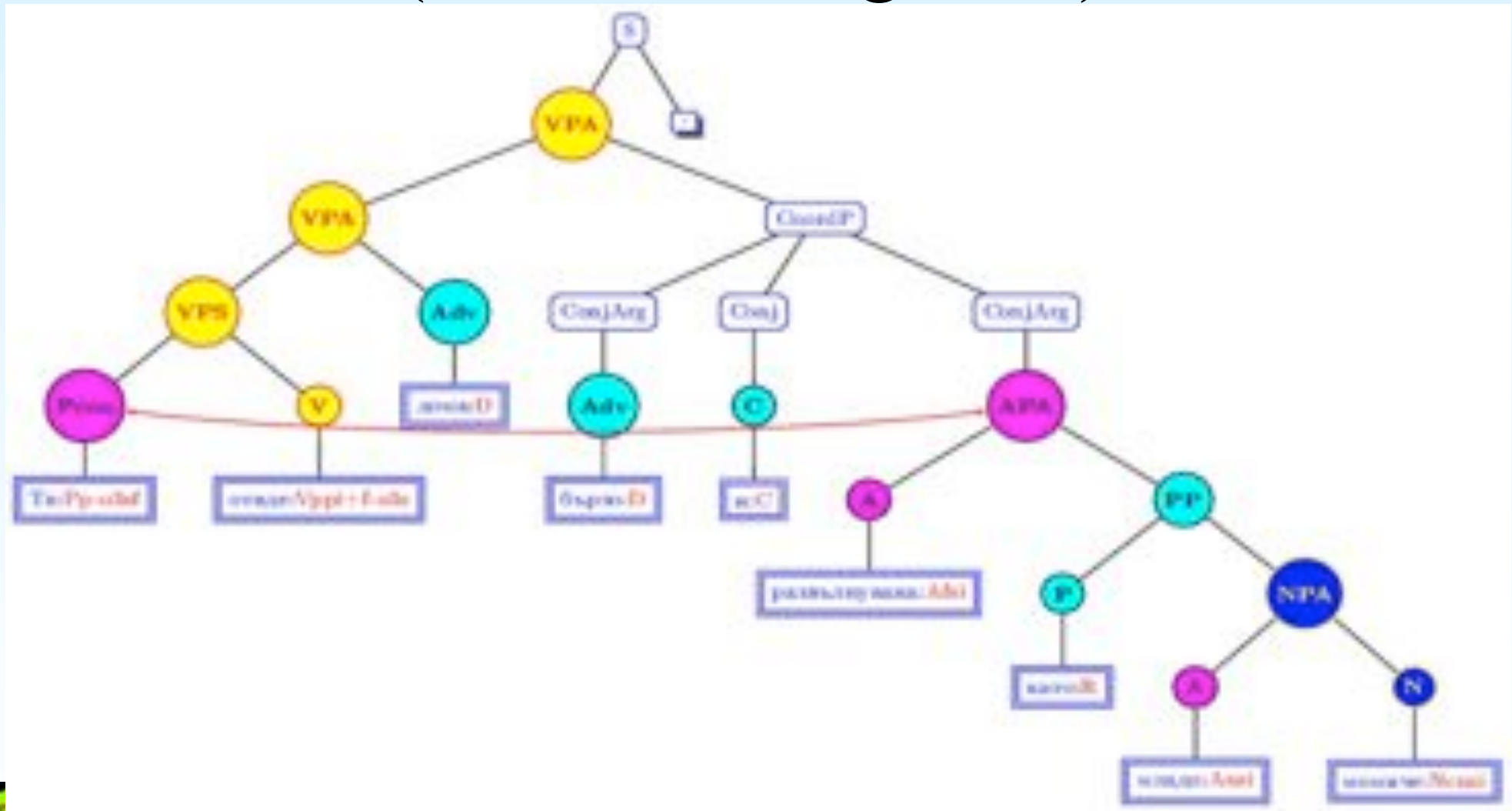


NP Internal



NP Coordination

Adjunct Coordination (Unlike Categories)



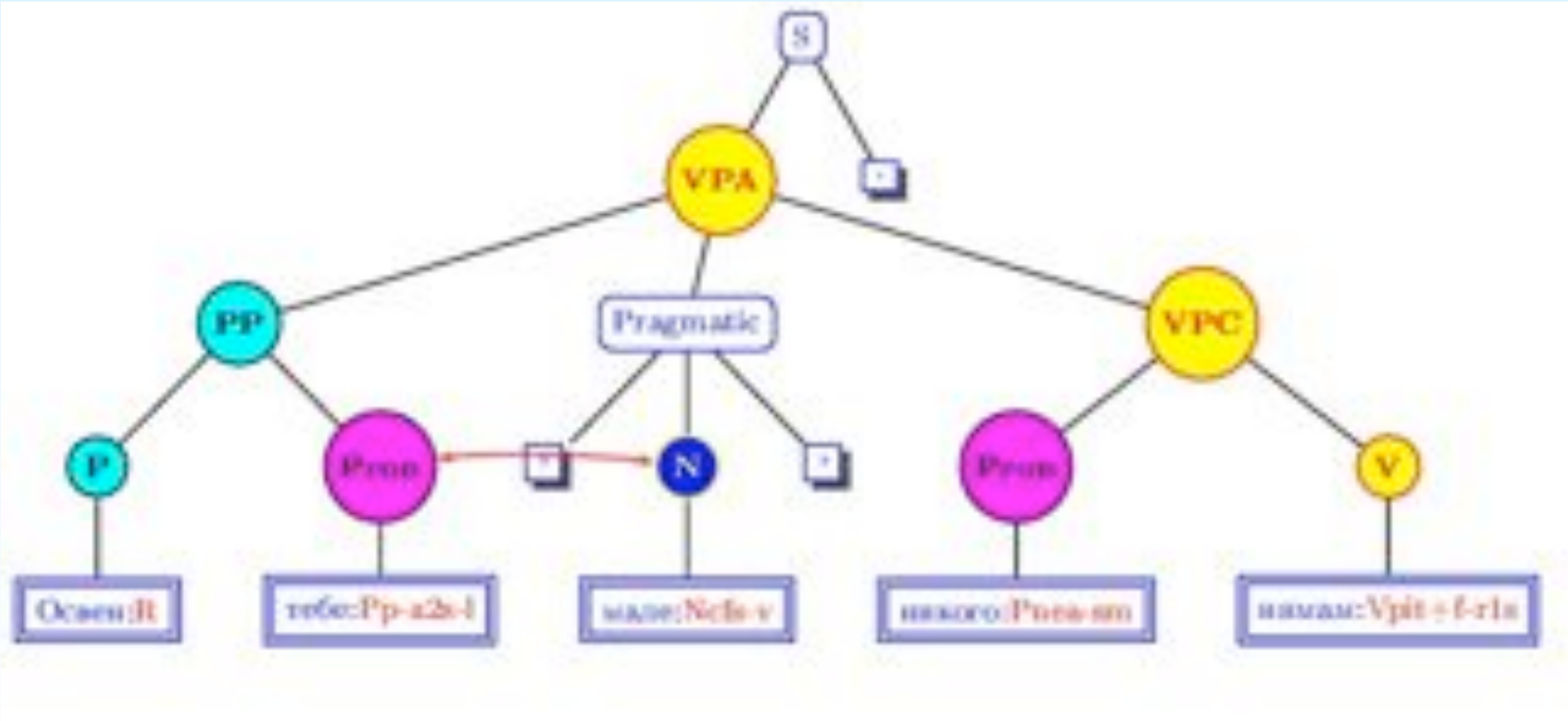
Pragmatic constituents

Elements of the sentences structure with primarily pragmatic impact

Here we include different kinds of parenthetical expressions (*of course*, *on the other hand*, etc), vocative phrases

They are attached to the phrases which they modify pragmatically as adjuncts

Pragmatic Adjunct



Core Phenomena (1)

Unexpressed Elements:

- Pro-dropness

  [kazah mu] [da prochete knigata]
  'I told him to read the book.'

- Ellipsis

  [Ivan pie bira,] [a Maria vino]
  'John drinks beer, but Maria wine.'

- Frame alternation

  [kazah mu] [da chete]
  'I told him to read.'

Core Phenomena (2)

- Co-referential Relations (**equality**, **member-of**, **subset-of**):
- Agreement
- Binding
- Anaphora resolution
- Definiteness
- Control

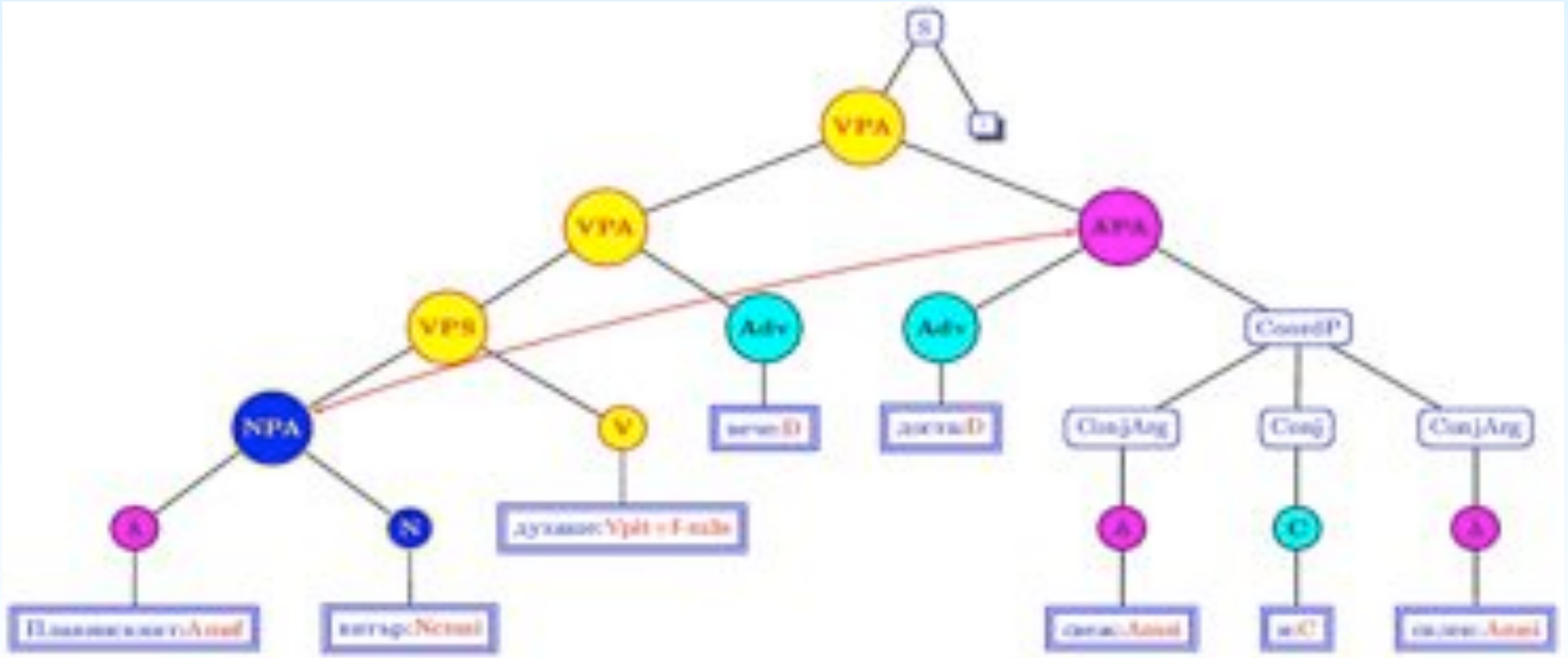
Core Phenomena (3)

- Relative clauses
- Secondary predication

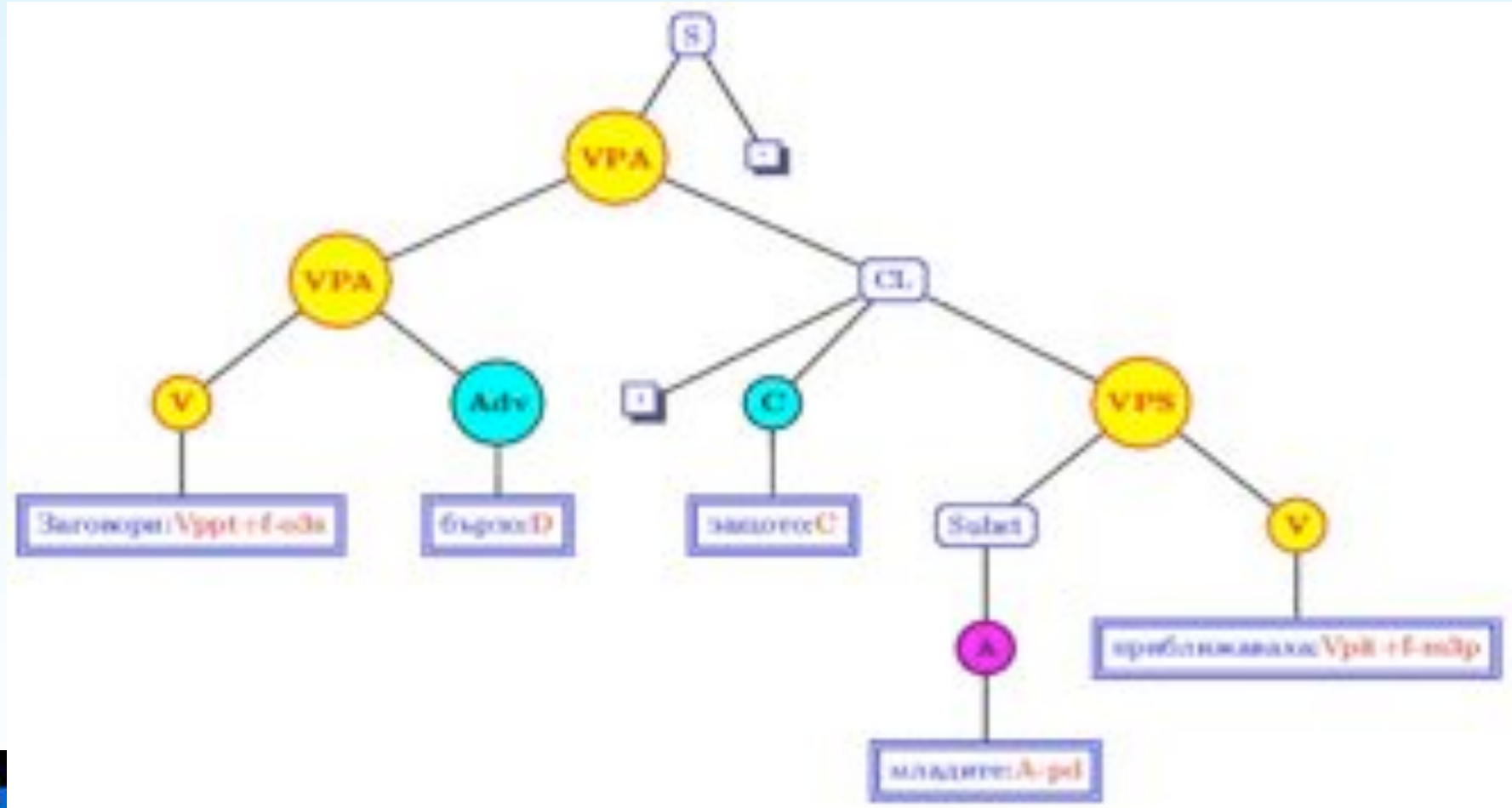
Type-shifting:

- Substantivization
- Nominalization
- Verbalization

Secondary predication



Substantivization



Summary BTB

- HPSG-based treebank
- Contains more than 15000 sentences
- Encodes dependency and constituent information
- Used in CoNLL 2006 task
- Good basis for development of manual and machine learning grammars of Bulgarian

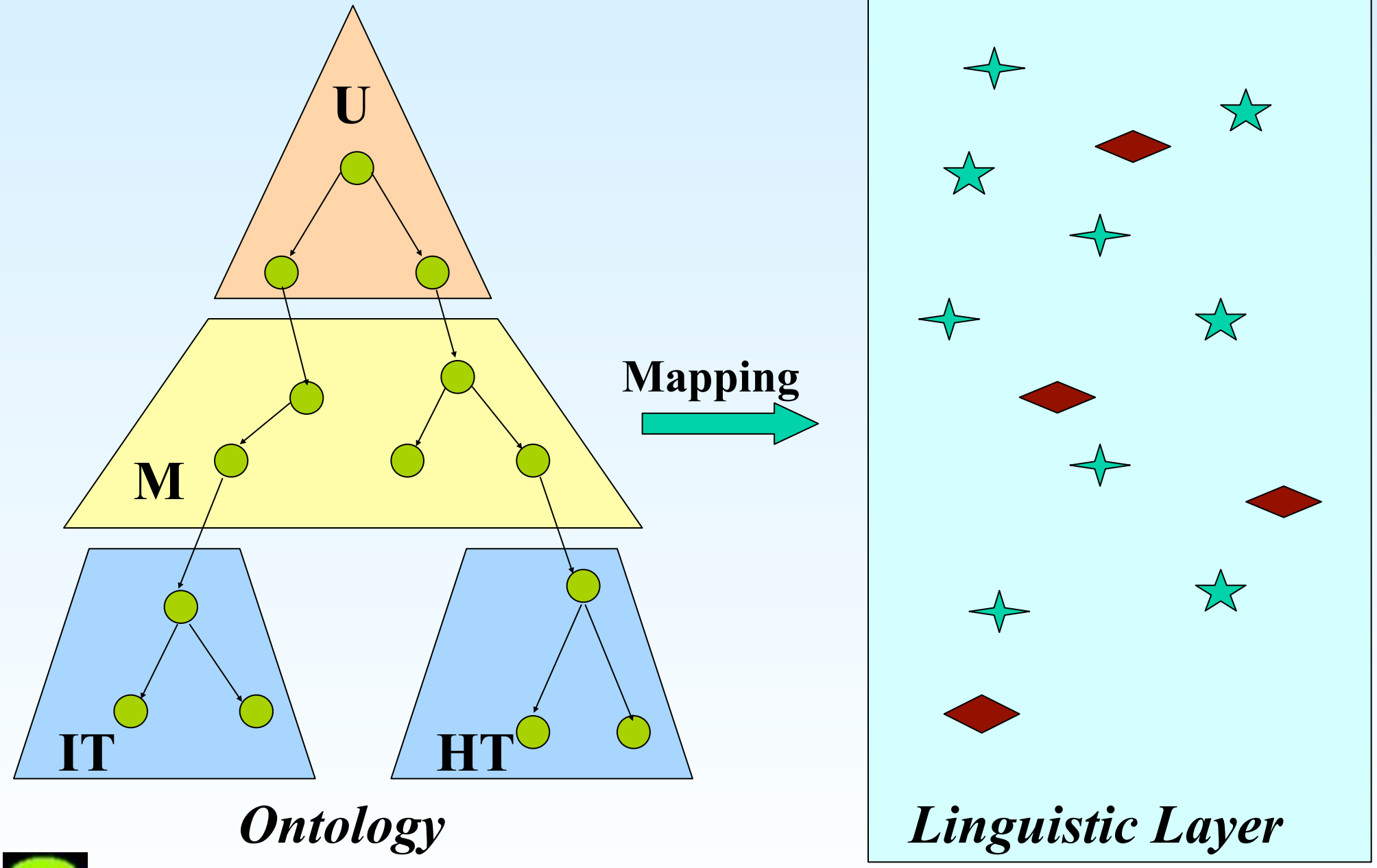
Ontology-based Lexicon

Based on Kiril Simov and Petya Osenova. 2008. Language Resources and Tools for Ontology-Based Semantic Annotation

Work done in projects: AsIsKnown, LT4eL, LTfLL

Ontology-Based Lexicon

- Conceptual part of the meaning is represented in a formal ontology
- Language specific part of the linguistic knowledge is encoded in the lexicon and the grammar of the language
- **Simultaneous construction of the formal ontology and the lexicon as part of ontology-to-text relation**


U
M
IT
HT
Mapping
Ontology
Linguistic Layer

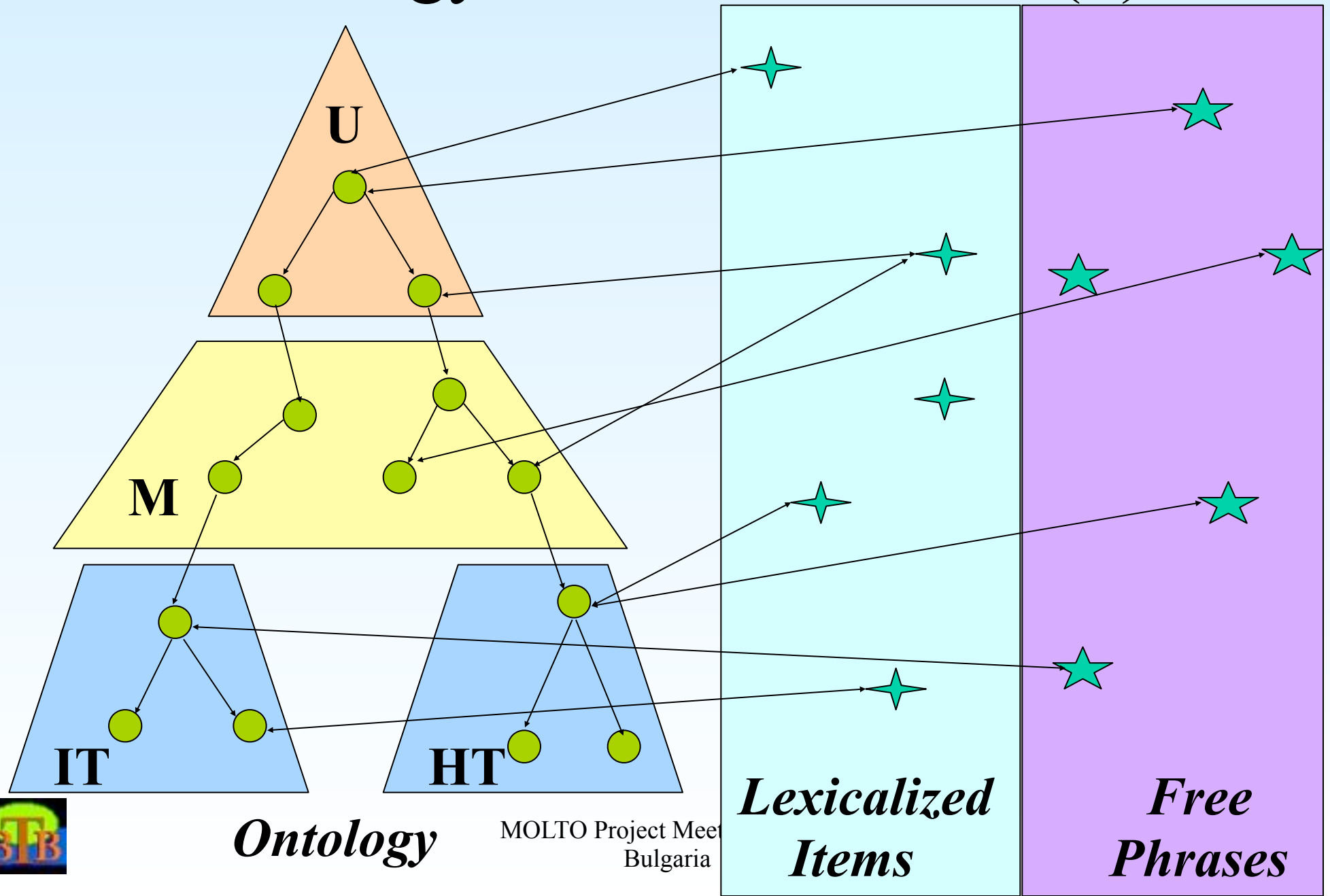
Lexicons and Annotation Grammars

- Creation of an instance of *ontology-to-text* relation
- Support the interaction with the users
- Support the semantic annotation

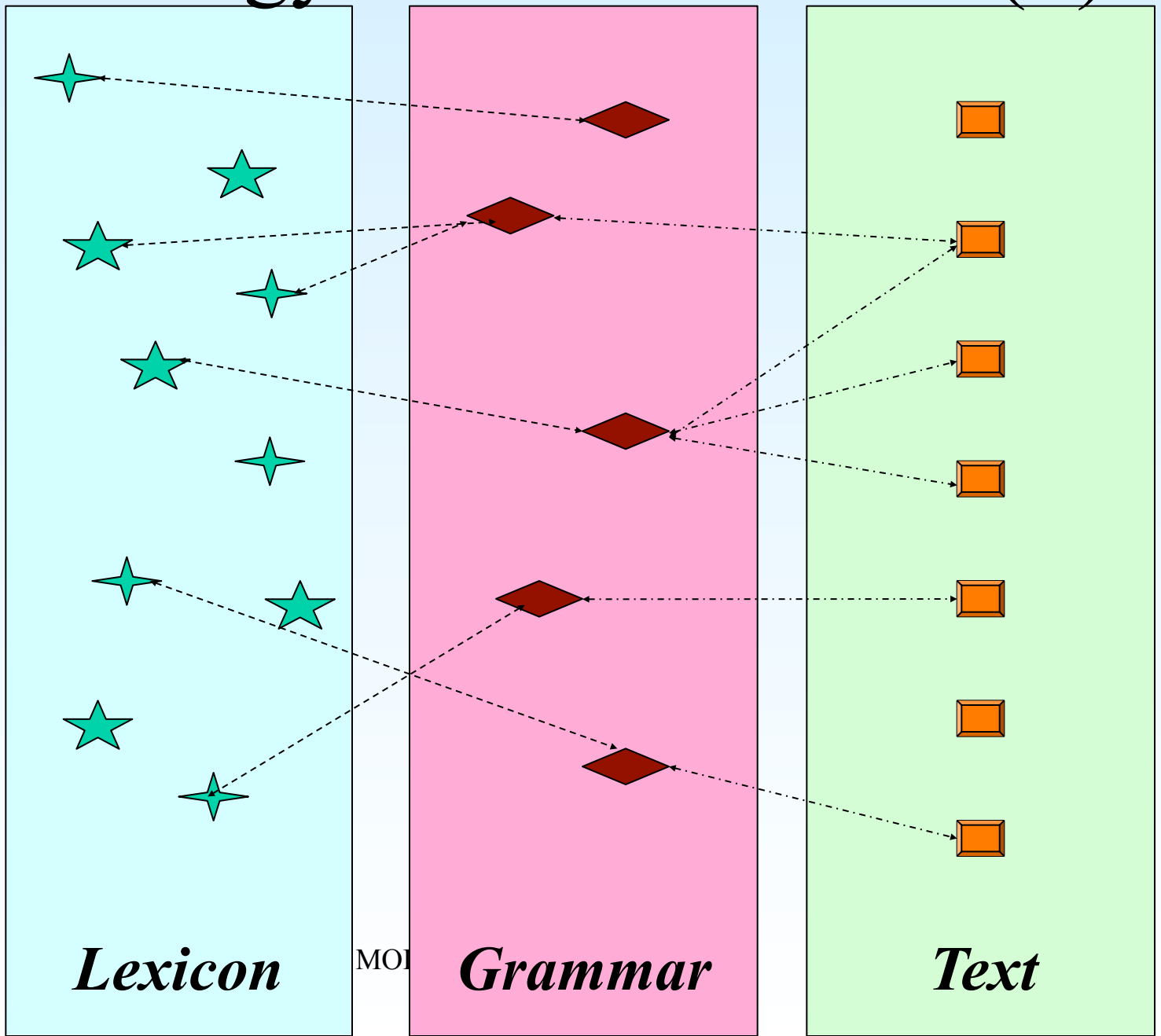
Ontology-to-Text Relation (1)

- Ontology is the repository for word senses
  - Polysemy and metonymy are encoded as interrelated concepts
- Lexicon represents the relation between word sense (concept, relation, instance in the ontology) and other lexical knowledge – morpho-syntactic features, etc
  - Human oriented features
- Grammar represents the relation between lexical items in the lexicon and their realization in the text

Ontology-to-Text Relation (2)



Ontology-to-Text Relation (3)



Roles of the Lexicon

- The lexicon interrelates the conceptual information from the ontology and the annotation grammar
- The lexicon is an interface between the user and the ontology
  - Navigation over ontology in the language of the user
  - Contextual variation – different lexicons for different users
  - Support creation of domain ontologies

Lexical Entry Structure

- Concept, relation or instance name
- List of terms expressing the corresponding conceptual entity
- Contextual information
- Grammatical features – link to the grammar
- Definition

Problematic Cases

- There is no a lexical unit for a concept in the ontology
  - We allow non-lexicalized phrases in the lexicon
  - We encourage the additions of such phrases in cases when there are lexicalized terms
- Important terms in the language miss appropriate concepts in the ontology
  - We extent the ontology in order to provide appropriate concept

Example from the Dutch Lexicon

```
<entry id="id60">
    <owl:Class rdf:about="lt4el:BarWithButtons">
        <rdfs:subClassOf>
                        <owl:Class rdf:about="lt4el:Window"/>
        </rdfs:subClassOf>
    </owl:Class>
    <def>A horizontal or vertical bar as a part of a window,
       that contains buttons, icons.</def>
    <termg lang="nl">
                <term shead="1">werkbalk</term>
                <term>balk</term>
                <term type="nonlex">balk met knoppen</term>
                <term>menubalk</term>
                <def> . . . </def>
    </termg>
 </entry>
```

Concept Annotation Grammar

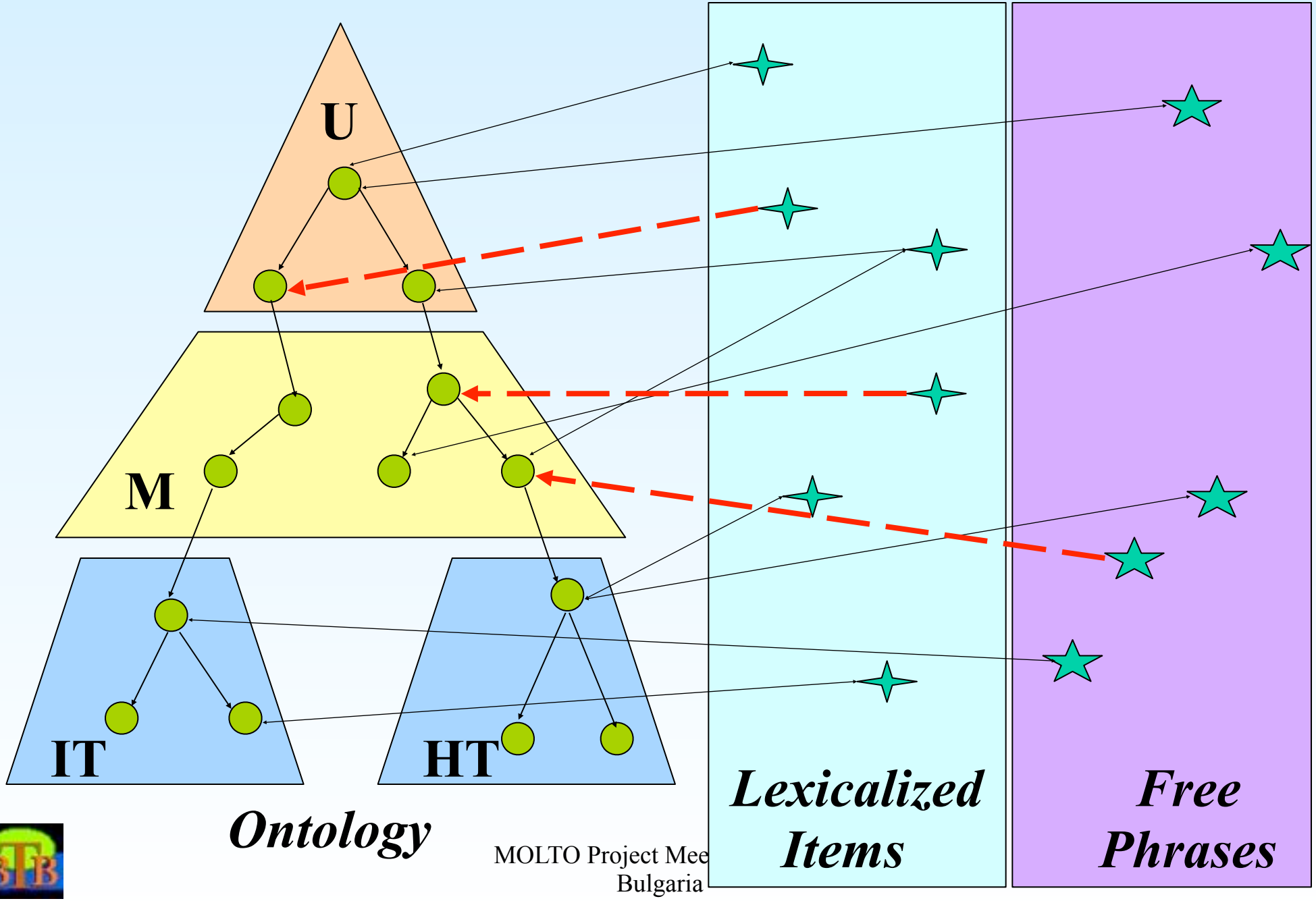
- Ideally, it is an extension of a deep grammar
- Minimally, it is a chunk grammar equipped with disambiguation rules for ambiguous terms
- The rules in the chunk grammar are created on the basis of the terms in the lexicon and rules from general chunk grammar
- Disambiguation rules are based on the local context and concept occurrences probability

Ontology-to-Text Relation (4)

- The *ontology-to-text* relation is a composition of the previous two relations
- It could support the following tasks:
  - Semantic annotation
  - Ontology-based search (including crosslingual search)
  - Ontology browsing
  - Ontology learning

Problems with the Model

- Both Lexicon and Ontology are artifacts – thus, not complete
- Lexicon is developed faster
- Ontology is constructed by extension of an Upper Ontology
- The ontology-to-text relation is defined by two relations: **equality** and **subsumption**


U
M
IT
HT
Ontology
Lexicalized Items
Free Phrases

Encoding of Valency

- Transferring the ideas from FrameNet and SIMPLE
- Ontology of Events – types and participants
- The lexicon maps the valency of lexical units to the participants encoded in the ontology (arguments or adjuncts)

Encoding of Metonymy

Of a special interest for semantic annotation are the ***metonymical*** and ***metaphorical*** uses of a lexical item

Definition of metonymy:

In general metonymy is defined as a trope in which one entity is used to stand for another associated entity

Examples: “stripe”

“*She was wearing stripe.*”

We represent ‘stripe’ as *Property* and thus it is connected to ‘textile’ via *property-of*. Then ‘textile’ which is *Material* and it is connected to ‘clothing’ again via the *used-for*.

The underlying meaning is: “*She was wearing a clothing made from a textile with a stripe design*.”

To Sum up on Metonymy

- Each metonymical usage introduces (at least) two semantic indices: one for the literal meaning of the word ('*stripe*' as a property) and one for the meant meaning ('*stripe*' as material for clothing – *material that has the property stripe*)
- In metonymic polysemy, both the basic and the secondary senses are literal.

Summary OBL

- Senses of lexical units correspond to concepts or relations in ontology
- Lexical relations are mapped to ontology relations
- Names are added as instances of concepts
- The remaining language knowledge is encoded in the lexicon and the grammar
- Ontology provides direct connection to world knowledge
- Ontology-to-text relation is grammar based

Bulgarian Language Technology

Based on Kiril Simov, Petya Osenova, Sia Kolkovska,

Elisaveta Balabanova, Dimitar Doikoff. 2004. *A Language Resources Infrastructure for Bulgarian*. LREC 2004, Lisbon, Portugal

Work done in project: BulTreeBank and CLaRK

Preliminary Notes (1)

A central question in HLT says:

"*what is minimally required to guarantee an adequate digital language infrastructure for a language*?"

Basic Language Resources Kit (BLARK)
is defined as a set of three groups:

- applications
- processing modules
- language data

Preliminary Notes (2)

The following questions are addressed:

- Language resources (LRs) with respect to the BLARK requirements
- The creation of a more advanced resource like a treebank and the basic language resources, which lack in this language
- The existent LRs as a solid basis for the development of other LRs

Treebanking as Basic Language Resources Compiler (1)

- The creation of a treebank for a “less-spoken” language like Bulgarian is a challenge
- The greatest problem -> the lack of a complete set of language resources
- Our decision: to produce a basic set of language resources for Bulgarian, which are easily adaptable for different mono- and multilingual NLP tasks

Treebanking as Basic Language Resources Compiler (2)

Two stages have been distinguished:

- Before starting the treebank creation – the implementation of basic processing modules
- Parallel to the treebank creation – the compilation of resources, which need more elaborate and high quality information

Treebanking as Basic Language Resources Compiler (3)

Two principles have been applied:

- **Bootstrapping principle -** its aim is to obtain as much information as possible at the very basic processing levels
- **Corpus-driven principle -** several results are simultaneously obtained by using extraction and observation procedures

BulTreeBank Language Technology

- Tokenizers - segmentation and classification
- Morphological analyzer (Lexicon more 110000 lemmas)
- Disambiguator(s)
- Partial grammars
  - sentence splitter
  - named-entity recognition module
  - chunkers

Morphological Analyzer

- Assigns all possible analyses to the tokens
- Implemented as a regular grammar
- Works together with the ‘token classification’ and with the gazetteers

Disambiguator(s)

- **Rule-based disambiguator** - a preliminary version of a rule-based morpho-syntactic disambiguator --> 80 % coverage
- **Neural-network-based disambiguator**

  Its accuracy is of 95.25 % for part-of-speech and 93.17 % for complete morpho-syntactic disambiguation

After the MorphoSyntactic Analysis and Disambiguation

<w aa="**Ncmsi**" ana="**Ncmsi**">**Човек**</w>

<w aa="**R**" ana="**R**">**с**</w>

<w aa="**Ncmsi**;**Vppt+cv--smi**" ana="**Ncmsi**">**опит**</w>

<w aa="**C**" ana="**C**">**и**</w>

<w aa="**Ansi**;**D**" ana="**Ansi**">**богато**</w>

<w aa="**Ansi**;**Ncnsi**;**Vpptcaosni**" ana="**Ncnsi**">**минало**</w>

Named Entity Recognition

Based on the information from the gazetteers and on Regular Grammar rules:

- numerical expressions
- names
- abbreviations
- special symbols

Example of Named Entity Annotation

<N>
<ph>Седем дни</ph>
<sort>OtherNE</sort>
<gramInt>Np-pi</gramInt>
<gramExt>Npfsi</gramExt>
<subsort>телевизия</subsort>
</N>

The Chunkers: General Assumptions

- Deals with non-recursive constituents
- Relies on *a clear-indicator* strategy
- Delays the attachment decisions
- Aims at accuracy, not coverage

Chunkers

- NP chunker
  - after preposition NPs
  - “sure” non-recursive NPs
- VP chunker
  - Analytical wordforms
  - “Da” constructions
  - Verb clitics
- PP, AP, AdvP, Clausal chunkers

After the Application of Some Chunk Grammars

- **Common NP chunks**
  [един човек] от [града] (‘one man from town-the’)
- **Name NP chunks: NEpers, NEloc etc.**
  [Министерство на културата] (‘Ministry of Culture’)
- **Complex NP chunks**
  [нашето [Министерство на културата]]
  (‘our Ministry of Culture’)
- **Analytical verb forms**
  [да [му я даде]] (‘to him her give-3p, sg’)
  to give it to him

Bulgarian Parsers

- Dependency parser within MaltParser trained on the treebank – more than 85 % accuracy
- BURGER – Bulgarian Resource Grammar based on Matrix Grammar implemented in LKB

HPSG-based Statistical Translation

- A new workpackage of existing European project EuroMatrixPlus started in July 2010
- The workpackage includes four tasks:
  - Creation of Bulgarian-English parallel HPSG treebank
  - Implementation of HPSG parser for Bulgarian
  - HPSG-based Statistical Translation Model
  - Evaluation

Bulgarian-English Parallel HPSG treebank

- Following the analyses of ERG and BURGER
- Aligned on several levels: sentence, word and structural
- Semiautomatic construction – BURGER, POS Tagger, MaltParser, Minimal Recursion Semantics
- Sources: BulTreeBank, ERG test sets, parallel corpora

HPSG parser for Bulgarian

- BURGER will be developed further to cover ERG data sets and BulTreeBank
- Combination of automatic modules to approximate HPSG parser for Bulgarian – POS tagger, dependency and constituent parsers trained over the treebank

HPSG-based Statistical Translation Model

- Two models:
  - Direct manipulation of feature structures, and
  - Incremental transfer on the basis of partial descriptions of feature structures
- Our goal is to use MRS representation as a basis for training of statistical transfer model
- In this work we will use an English to Bulgarian lexicon and the Bulgarian Ontology-based Lexicon

Conclusion

- Bulgarian has enough language resources and technology to support language applications
- But more work is necessary to make these resources compatible, freely available (at least for research)
- More resources are necessary for the semantic processing, speech processing and multimodal applications