BulTreeBank
Bulgarian Language Resources and Technology

Kiril Simov and **Petya Osenova**

Linguistic Modelling Department
Institute for Information and Communication Technology
Bulgarian Academy of Sciences
AComin Project

CLTL, VU University Amsterdam
9 July, 2014

Plan of the Talk

- AComin Project
- BulTreeBank Group Language Resources and Tools
  - Bulgarian HPSG-based Treebank
  - Valency Lexicon
  - Ontology-to-Text Relation Approach (Ontologies)
  - Bulgarian Language Technology
- Current Projects

Whom to thank for being here?

AComIn: Advanced Computing in Innovation

- A 3,2 M€ grant in FP7 Capacity with a single beneficiary – IICT-BAS
- Objectives:
  - Strengthening Human Potential
  - Providing up-to-date Research Infrastructure
  - Focus on users
  - Networking with EU partners
  - Strengthening the IICT-BAS Innovation Capacity
  - Dissemination via various events/channels
  - Organising assessment of the IICT-BAS achievements

Human Potential

- A major budget
- Employment of Incoming Experienced Researchers
- Identified areas of IICT expertise with available critical mass of researchers:
  - *Advanced computing*
  - *Smart interfaces*
  - *Optimization and intelligent control*
  - *Language and semantic technologies*
- AComIn considers 2 categories of incoming staff:
  - Post-docs for long-term employment
  - Experienced researchers with more than 10 years of experience for shorter employments

Purchasing Smart Lab and Building User Communities

- A coherent set of complementary high-tech devices: 3D input/output of microstructures, modelling of particle dynamics and interaction in granulated media as well as processing of speech, images and signals.
- Ensure 'data autonomy' of IICT-BAS
- Viewed as advanced 'periphery'
- Enable technologies for processing of microstructures and dynamic events – central topics in H2020
- Support synergy of IICT-BAS research and upgrade to modern computational paradigms
- Unique set of devices for South-East Europe

BulTreeBank Group – LT Synopsis (1)

- **BulTreeBank** – An HPSG-based treebank of Bulgarian.
- **Bulgarian Reference Corpus BulTreeBank** – Texts annotated up to paragraph level with respect to TEI guidelines (near 400 million words)
- **BulTreeBank Morphosyntactic Corpus** – Annotated with grammatical information
- **English-Bulgarian Parallel corpora**
- **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
- **Domain ontologies** in IT and Textile domain

BulTreeBank Group – LT Synopsis (2)

- **WebCLaRK** portal to resources and services (www.webclarg.org): Customizations for political speech: political.webclark.org; Old Bulgarian (to be released soon).
- **BulTreeBank Partial Grammar**
- **Dependency Parser for Bulgarian**
- **Morphological taggers** (Morpho-lexicon-based; Neuron-network-based; Guided Learning-based)
- **HPSG-based** resource grammar for Bulgarian

BulTreeBank (BTB)

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 (2001-2004; 2005-2007)

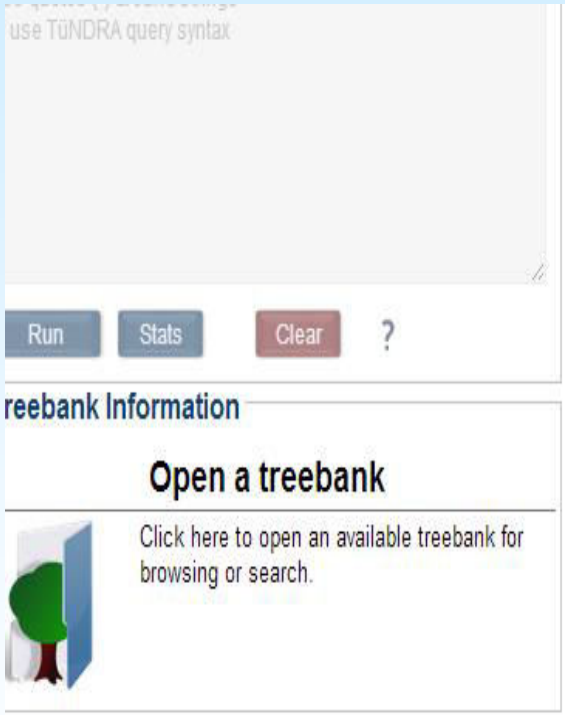
Funded by: Volkswagen Foundation, Germany

Formats and Visibility of BTB

- HPSG-based treebank – 215 000 tokens
- BulTreeBank-DP – 196 000 tokens (without ellipsis) (http://www.bultreebank.org/dpbtb)
- Developed with an in-house system: CLaRK

- Searchable in treebank web services **INESS** (http://clarino.uib.no/iness/main-page) and **TUNDRA** (https://weblicht.sfs.uni-tuebingen.de/Tundra/)

https://weblicht.sfs.uni-tuebingen.de/Tundra/



Select a treebank

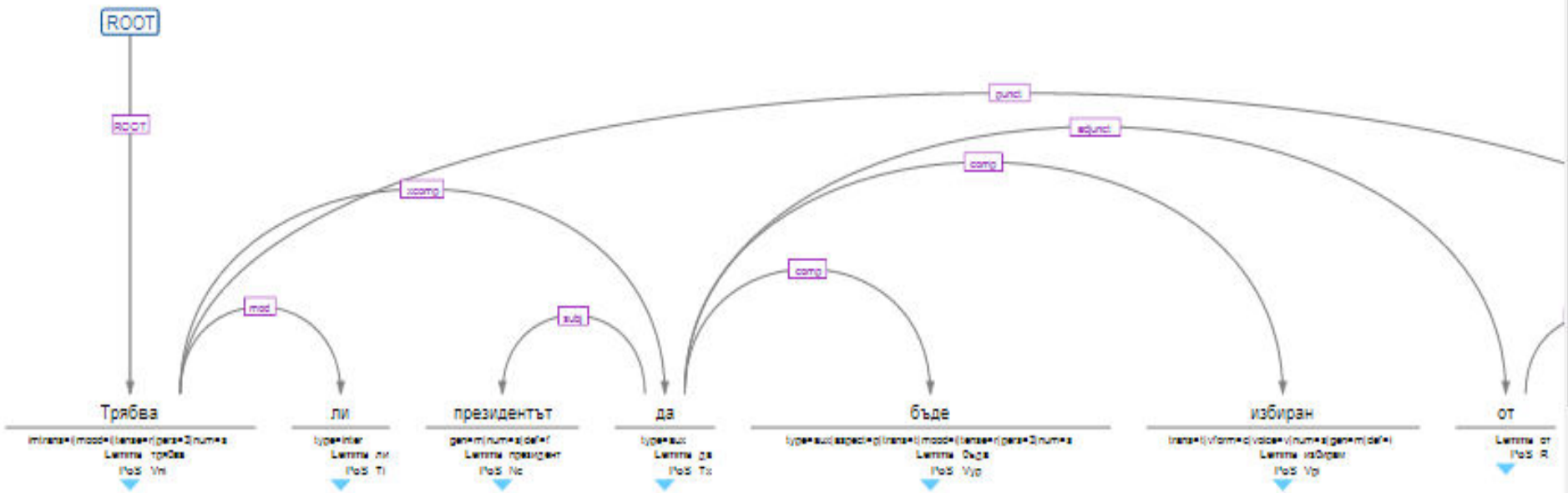
	name	type	
	**TüBa-D/DC: Der Schatz im Silbersee, Karl May**	Constituency	Automatically parsed edition of "Der Schatz im Silbersee" (1894) by Karl May, from the TüBa-D/DC.
	HPSG-based Syntactic Treebank of Bulgarian	Dependency	The main objective of BulTreeBank Project was to create a high quality set of syntactic structures of Bulgarian sentences within the framework of HPSG. Ideally, the tree bank should contain samples of all the syntactic structures of the language.

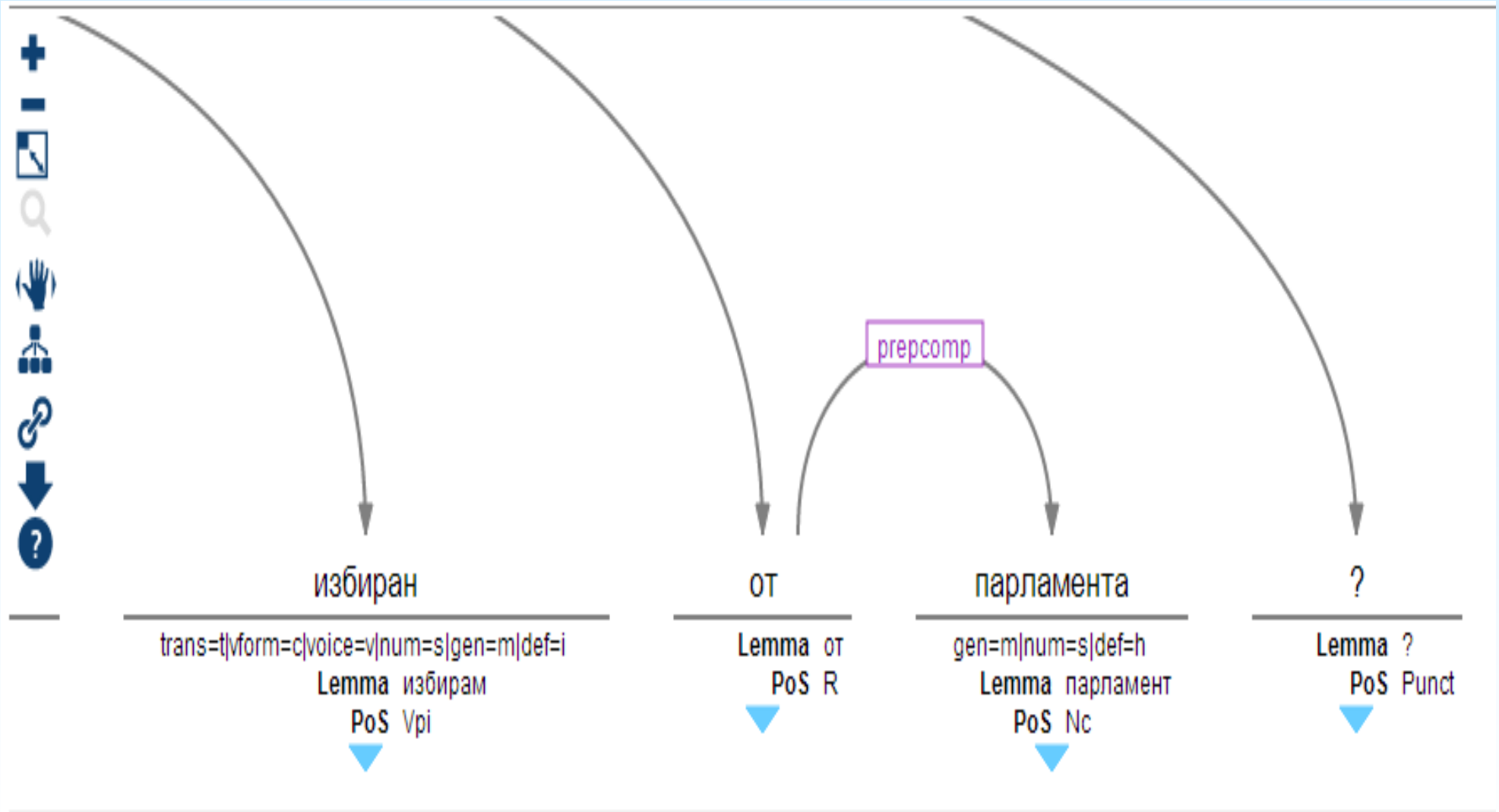
Tree

Browse Treebank

Tree No.: 4713 / 11990

#4713: Трябва ли президентът да бъде избиран от парламента?




prepcomp
избиран
trans=t|vform=c|voice=v|num=s|gen=m|def=i
Lemma избирам
PoS Vpi
от
Lemma от
PoS R
парламента
gen=m|num=s|def=h
Lemma парламент
PoS Nc
?
Lemma ?
PoS Punct

INESS :: Home

clarino.uib.no/iness/main-page

Apps | Word to PDF Online ... | Clark web system | web.jhu.edu/bin/g/i... | Bonnie J. Dorr's Publ... | Оперативна програ... | aclweb.org/antholo... | Елка Стефан

iness

Sign in: Local | eduGAIN | OpenIdP

Welcome to INESS

- Main Page
- Project description
- Participants
- Documentation
- Publications
- Links
- Resources

Treebanks

- Treebank selection
- Parallel Sentences
- Metadata

Tools

- XLE-Web

Text

INESS is an open system serving a range of research needs, offering an interactive, language independent platform for building, ac searching and visualizing treebanks. All its functionality can be used online through a browser.

INESS hosts treebanks for many languages. We are building a large treebank for Norwegian, obtained by parsing automatically with grammar. Part of the corpus will be efficiently manually disambiguated with the LFG Parsebanker, while the rest will be stochastica disambiguated.

Access to most treebanks is restricted to registered users after login.
If you are connected to a Norwegian institution of higher education (at present UiB and UiO), you can log in via Feide (please use *eduGAIN* link). If you have an identity provider (IdP) that is a member of eduGAIN, chances are that you will be able to log in via y (please use the *eduGAIN* link). Otherwise, you will have to register an OpenIdP account. After you have logged in at INESS for the time, you will be given access to treebanks, depending on your needs.

Please contact us if you want to use INESS for your treebank.

The INESS treebanking environment

INESS offers an interactive, language independent system for *building, accessing and exploiting* treebanks

Choose a treebank:

- **Treebank selection:** Select treebanks according to specific criteria such as language, grammar type and project of origin
- Choose between **monolingual** and **parallel** treebanks

Sign in: Local | eduGAIN | Open

Treebanks

Choose a set of treebanks to work with. ?

Languages: All · Ancient Greek (to 1453) (4) · **Bulgarian** (1) · Church Slavic (3) · Classical Armenian (2) · English (1) · Georgian (2) · German (6) · Gothic (1) · Hungarian (1) · Icelandic (1) · Indonesian (1) · Latin (5) · Northern Sami (15) · Norwegian Bokmål (1) · Old English (ca. 450-1100) (5) · Old French (842-ca. 1400) (5) · Old Norse (6) · Polish (2) · Portuguese (4) · Spanish (10) · Turkish (1) · Urdu (1) · Wolof (1)

Treebank Collections: **All** · **BulTreeBank** (1) · GEGO (0/2) · HunGram (0/1) · ISWOC (0/23) · IcePaHC (0/1) · Menotec (0/6) · Mercurius (0/1) · NorGram (0/1) · POLFIE (0/2) · PROIEL (0/15) · ParGram (0/12) · Sami-open (0/15) · TiGer (0/3) · WolGram (0/1)

Treebank Types: **All** · lfg (0/14) · constituency (0/5) · dependency (0/6) · **dependency-cg** (1/16) · dependency-proiel (0/38)

☐ Show only Parallel Treebanks

Reset

Distribution info

Availability: available-restrictedUse

Availability start date: 2012-03-29

Licence info

Licence: MS-NC-NoReD

By accepting the terms of the license you will be granted access to the resource.

Accept

Restrictions of use: academic-nonCommercialUse, attribution, noRedistribution

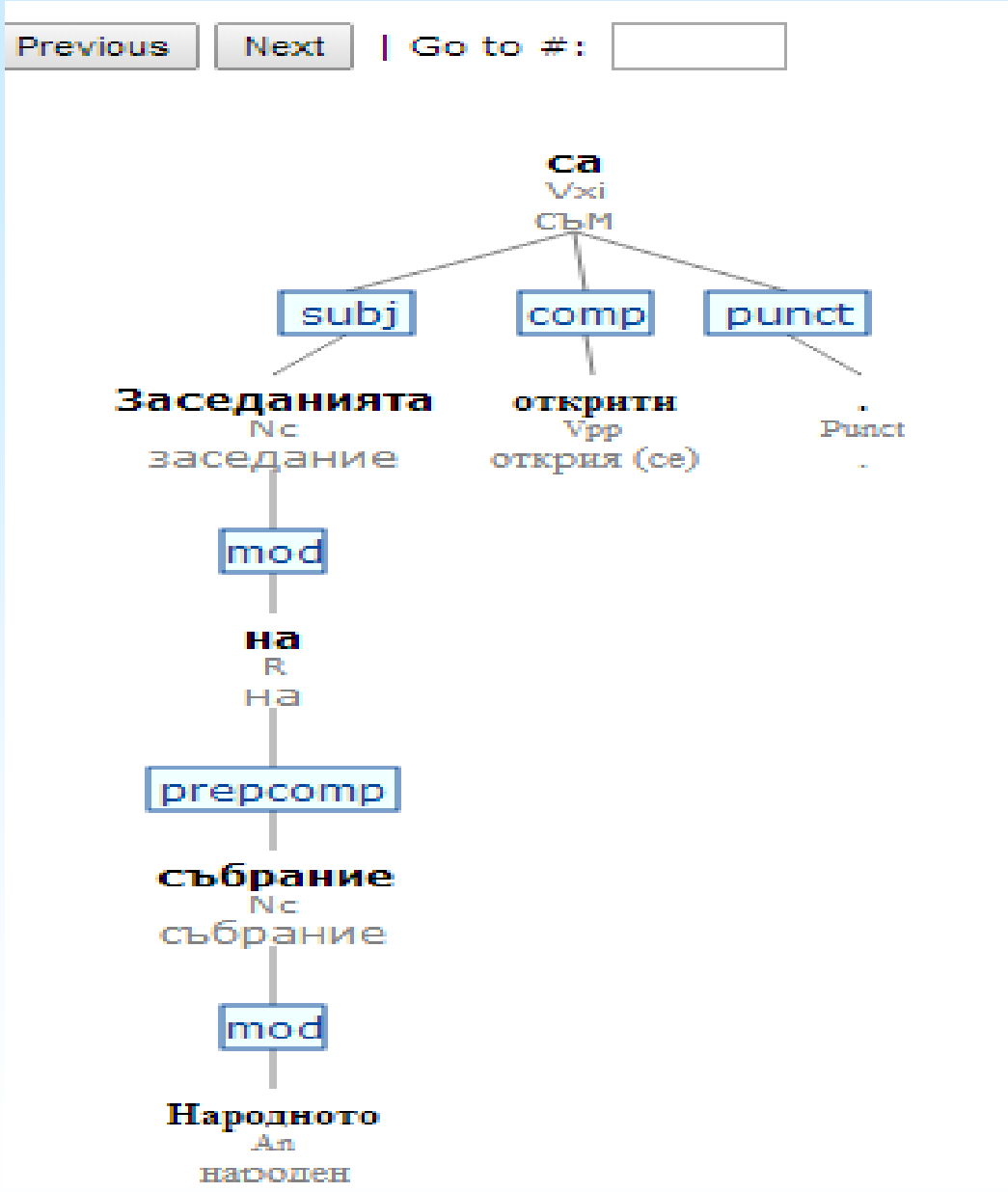
Distribution access medium: downloadable

Download location: http://www.bultreebank.org/dpbtb/, http://iness.uib.no

User nature: academic

50 sentences chosen. Previous Next | Rows per page: 50

Id	Uid	Words	Sentence
1844	2139060	9	Народното събрание осъществява законодателната власт и упражнява парламентарен контрол.
1845	2139061	8	Народното събрание се състои от 240 народни представители.
1846	2139062	9	Народното събрание се избира за срок от четири години.
1847	2139063	29	За народен представител може да бъде избиран български гражданин, който няма друго гражданство, навършил е 21 години, не е поставен под запрещение и не изтърпява наказание лишаване от свобода.
1848	2139064	14	Кандидатите за народни представители, които заемат държавна служба, прекъсват изпълнението й след регистрацията си.
1849	2139065	15	Законността на изборите може да се оспори пред Конституционния съд по реда, определен със закон.
1850	2139066	6	Обвързването със задължителен мандат е недействително.
1851	2139067	16	Народните представители действат въз основа на Конституцията и законите в съответствие със своята съвест и убеждения.
1852	2139068	23	Народните представители не могат да изпълняват друга държавна служба или да извършват дейност, която според закона е несъвместима с положението на народен представител.


Previous
Next
| Go to #:
са
Vxi
съм
subj
comp
punct
Заседанията
Nc
заседание
открити
Vpp
открия (се)
.
Punct
.
mod
на
R
на
prepcomp
събрание
Nc
събрание
mod
Народното
An
народен

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

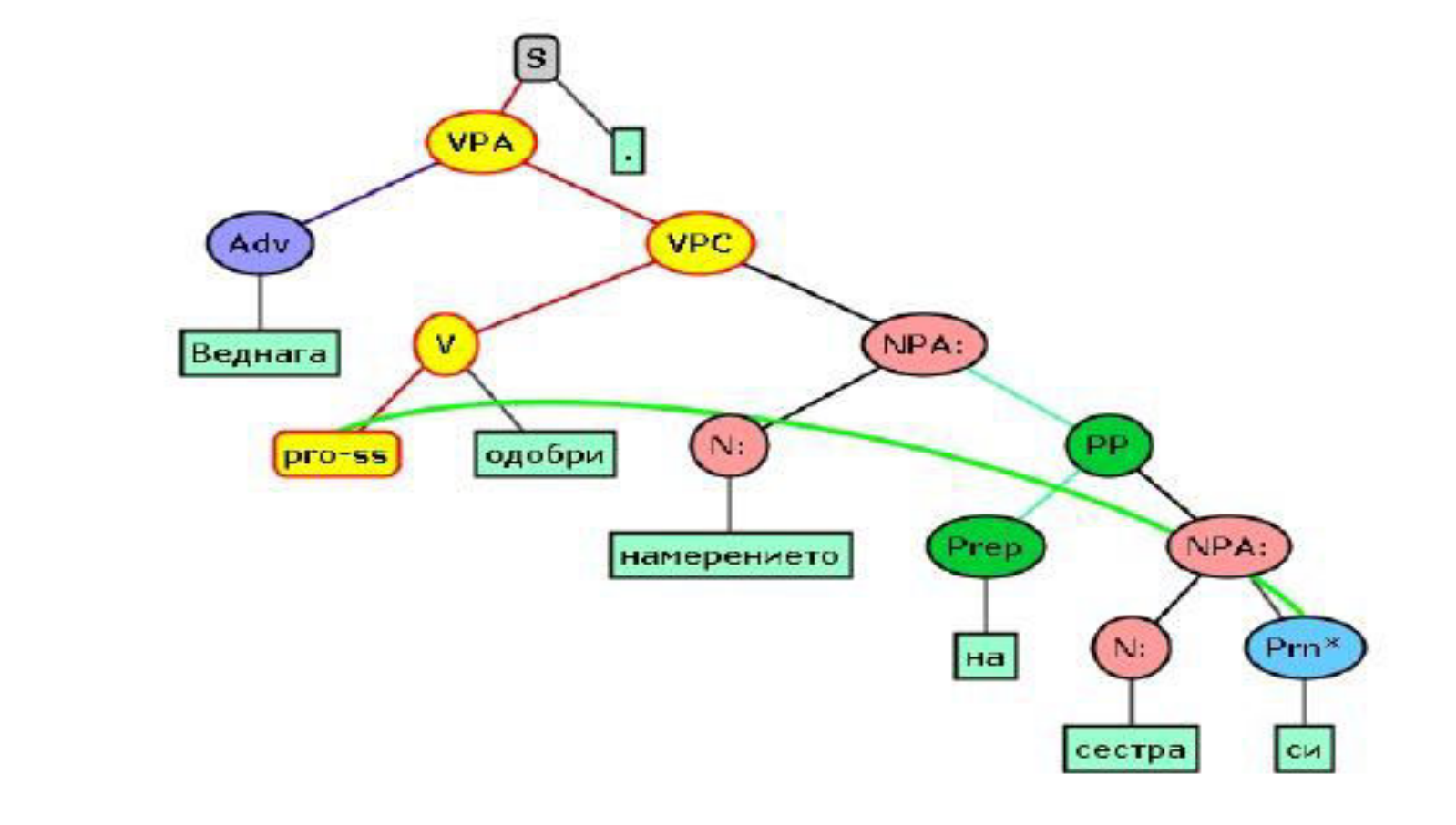
Core Phenomena (3)

- Relative clauses
- Secondary predication

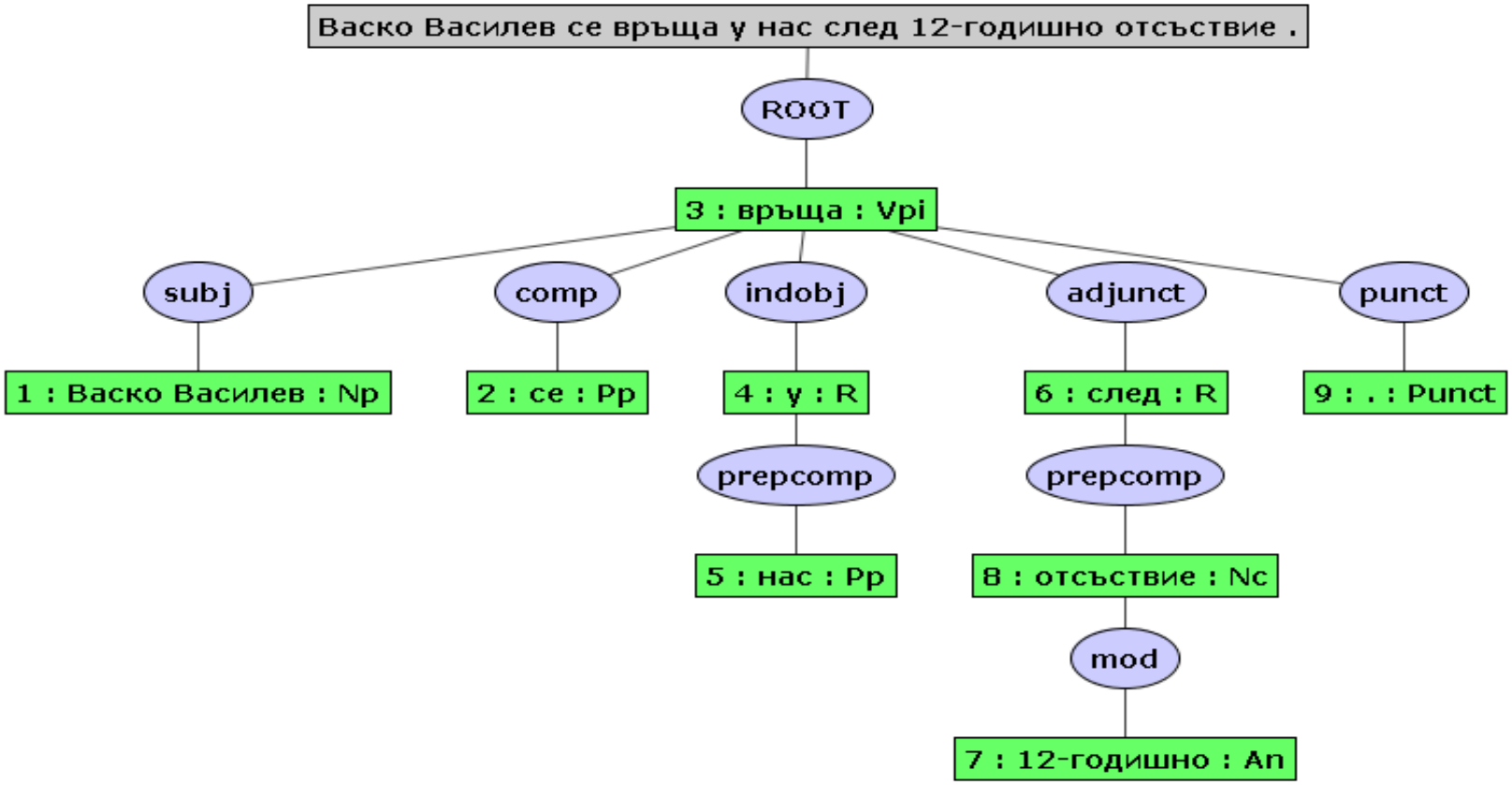
Type-shifting:

- Substantivization
- Nominalization
- Verbalization

HPSG-based representation



Dependency Representation (CoNLL-based)



Current developments (1)

- **Semantic Annotation** (80% completed; to be checked for consistency; to complete the missing relations to WordNet)
- **Valency Lexicon** (extracted; to be checked for consistency; to be finegrained for its mappings to DOLCE ontology)
- **Extending the BulTreeBank-DP** with more sentences
- **Handling ellipses** in the Dependency Part

Current developments (2)

- Creation of an English-Bulgarian Treebank
- Two approaches:
  - Processed text by English and Bulgarian Resource grammars . Problems: Bulgarian grammar does not have enough coverage)
  - Enhancing aligned words and phrased with (R)MRS analyses. Problems: propagated errors from the dependency parser and under or over-generation of the projecting rules.

Semantic Annotation

- Coverage of: Nouns, Adjectives, Verbs, Adverbs
- Mappings to the core/base WordNet concepts. In cases where they are not available, exploration of a pre-curated Bulgarian Explanatory Dictionary

Some Statistics

	tok	lemma	def		Tok/Lemma	Def/Lemma
Adj	17741	2077	4344		8.541646606	2.091478093
Adv	7571	533	726		14.20450281	1.362101313
Noun	49658	3477	8046		14.28185217	2.314063848
Verb	17977	2058	5163		8.735179786	2.508746356
Total	92947	8145	18279			
Tok/Lemma	11.41154082					
Def/Lemma	2.244198895					

Semantic Annotation

tok: select : виолетов : виолетова
tok: select : виртуозен : виртуозния
tok: select : вирусен : вирусни
tok: select : висок : висок
висок
S::Тя никога не се смееше с @@@ висок @@@ глас .
def : възвишен
@@@Selected@@@ : За звук, глас - който се произвежда от
def : за звук, глас - който се произвежда от колебания,
def : който има голям размер от долу нагоре по отвесна
def : Който има по-голям от нормалния размер, степен, и
def : Който има по-голяма от средната за даден вид пред
def : който превъзхожда другите; отличен
def : отдалечен от земната повърхност или издигнат над
def : твърде силен, много голям

Verb Valency Dictionary

- Covers **3283** lemmas in BulTreeBank
- The number of distinct valence frames for these lemmas is **6469**
- The average is about **2 valence frames** per lemma

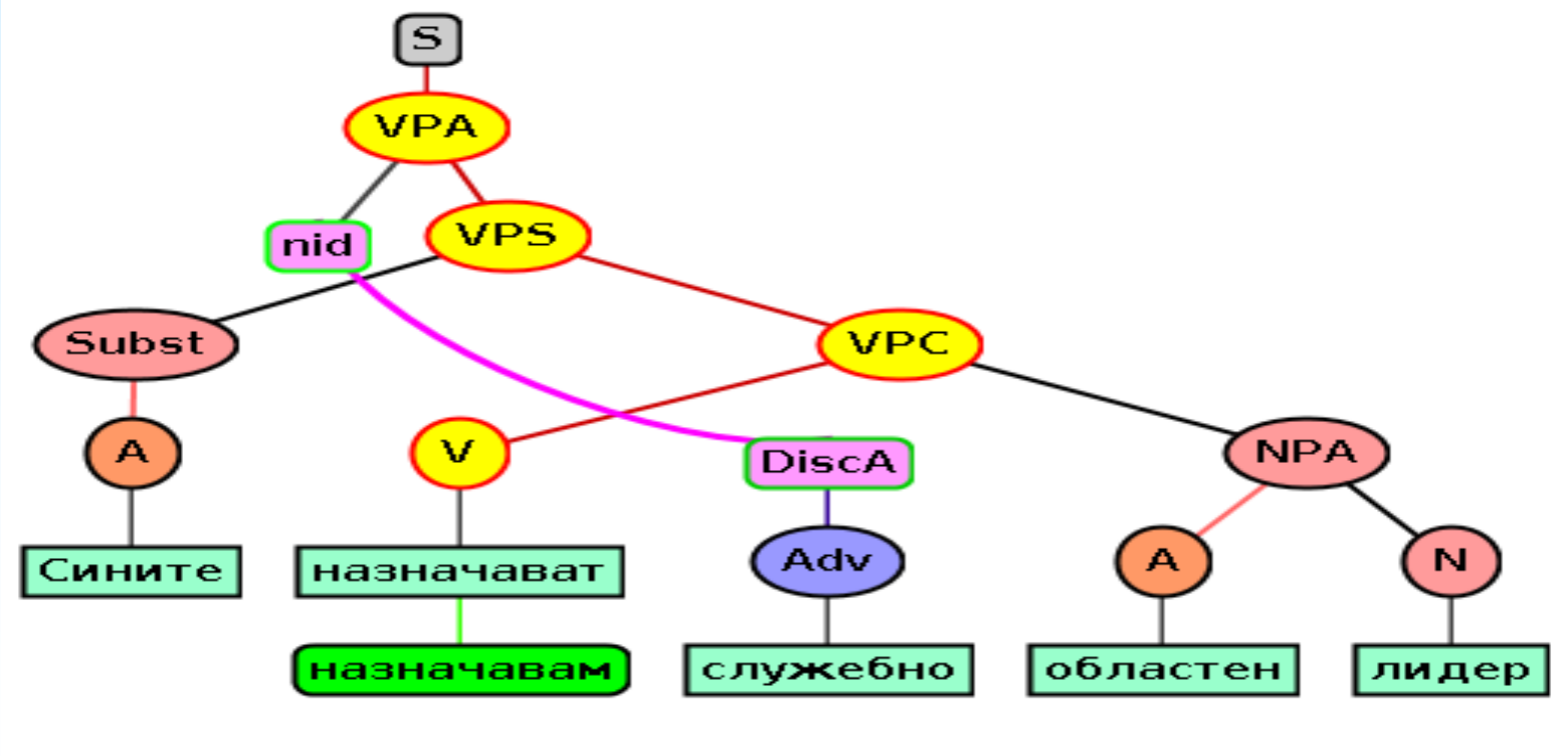
OntoValence Lexicon Extraction and Manipulation

- All the verbs have been extracted together with the sentences they have been used in
- Then they have been lemmatized and sorted by the lemma marker
- A default valence frame has been inserted, which presents a predicate with a **SUBJ**, **DIROBJ** and **INDOBJ**

Original representation of a sentence tree

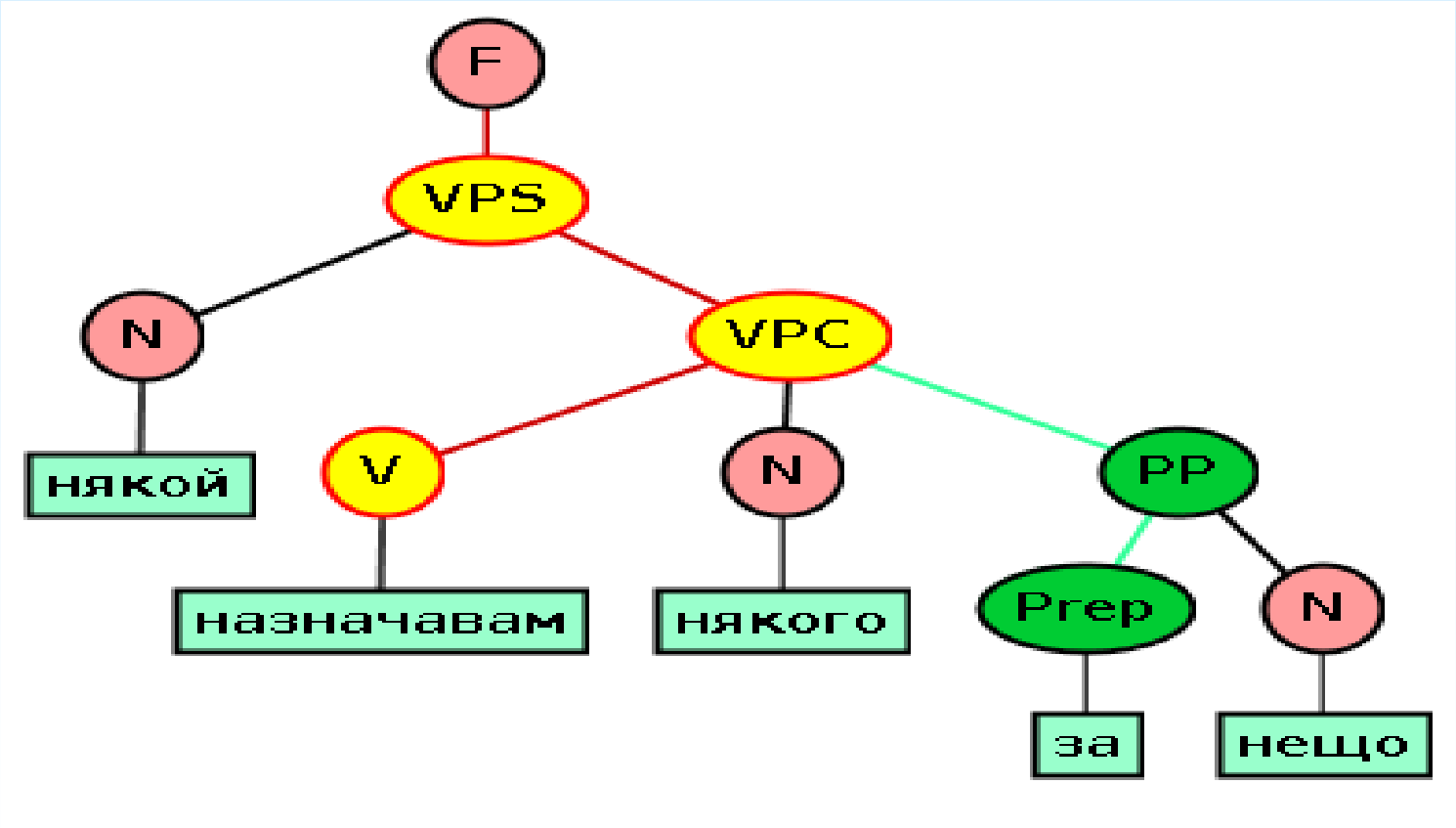
Gloss: *Blue-the appoint officially area leader*.

Translation: *The blue team ex officio appoints an area leader*.



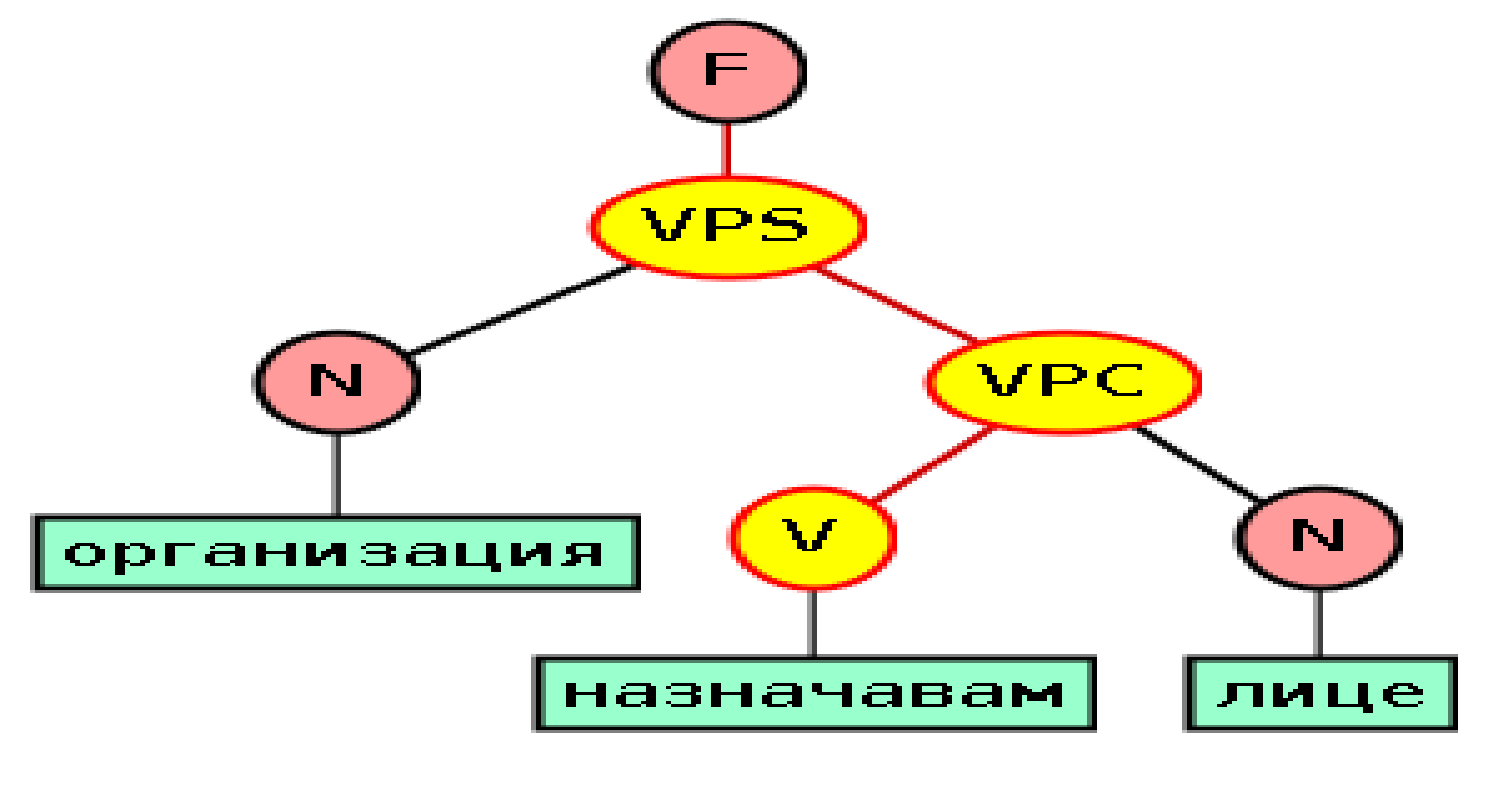
Default inserted tree

[SOMEBODY appoints SOMEONE for SOMETHING]



Resulting Frame

ORGANIZATION [appoint - lemma] PERSON



Some statistics

- The extracted annotated frames from BulTreeBank are **18081**
- Additional example material has been extracted also from the Bulgarian National Reference Corpus (when examples **< 5**)
- In BulTreeBank:
  - **920 verb lemmas** have occurred in only **once**;
  - **313 lemmas** have occurred **2 times**;
  - **200 lemmas – 3 times**;
  - **115 lemmas** – **4 times**;
  - **94 lemmas** – **5 times**

OntoValence Lexicon Architecture and Principles

Label	Description
VPA	head (verb)-adjunct
VPC	head(verb)-complement
VPS	head(verb)-subject
NPA	head(noun)-adjunct
NPC	head(noun)-complement
PP	head(preposition)-complement
PPA	head(preposition)-adjunct
APC	head(adjective)-complement
APA	head(adjective)-adjunct
AdvC	head(adverb)-complement
AdvA	head(adverb)-adjunct

Table 1: Description of the syntactic labels in BulTreeBank

Specifics

- The valence frame is kept to the surface syntax
- Thus, the *pro-drops* of any kinds are also presented within the frames
- The frame considers the *clausal complements* as well
- We encode the verb usage in *active voice*
- The verbs in *perfective and imperfective aspect* are considered separate lemmas
- The frame includes only the *inner participants* (semantically obligatory for the event or situation, presented by the predicate, but might be unexpressed on the surface level). We follow (Pustejovsky 1998)

Some Observations over the current state of the OntoValence Lexicon

N	Syntactic Frame Type	Number of Frame Occurrences
1.	Predicate – Direct Object (NP)	4089
2.	Subject (NP) – Predicate – Direct Object (NP)	3122
3.	Subject (NP) – Predicate	1339
4.	Subject (NP) – Predicate – Indirect Object (PP)	1243
5.	Predicate	1082
6.	Predicate – Direct Object (NP) – Indirect Object (PP)	1013
7.	Predicate – Indirect Object (PP)	888
8.	Predicate – Complement (CLDA)	807
9.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	695
10.	Subject (NP) - Predicate – Complement (CLDA)	643

Table 2: Frequency of syntactic Frames

Ontological Types:

EVENT > PERSON > OBJECT > ARTEFACT > COGNITIVE FACT

N	Syntactic Frame	Ontological Label
1.	Predicate	No Ontological Restrictions
2.	Predicate – Complement (CLDA)	EVENT
3.	Subject (NP) – Predicate	PERSON
4.	Predicate – Direct Object (NP)	PERSON
5.	Subject (NP) - Predicate – Complement (CLDA)	PERSON - EVENT
6.	Predicate – Direct Object (NP)	OBJECT
7.	Subject (NP) - Predicate – Direct Object (NP) – Indirect Object (PP)	PERSON – ARTEFACT – (for) OBJECT
8.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - PERSON
9.	Predicate – Direct Object (NP)	COGNITIVE FACT
10.	Subject (NP) – Predicate – Direct Object (NP)	PERSON - OBJECT

Main challenges

- The main problems in inter-annotator agreement seem to be:
  - the granularity of concept mappings as well as
  - the metaphorical and idiomatic usages

Other publications

- Osenova and Simov. Bulgarian Treebank Stylebook - http://bultreebank.org/TechRep/BTB-TR05.pdf (appeared as a book in Bulgarian in 2007) http://www.bultreebank.org/bgpapers/FormalGrammarBG.pdf
- Osenova et. al 2012: Petya Osenova and Kiril Simov and Laska Laskova and Stanislava Kancheva. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. LREC 2012 Proceedings.
- Simov and Osenova 2012: Kiril Simov and Petya Osenova. Bulgarian-English Treebank: Design and Implementation. In*: Linguistic Issues in Language Technology*, vol. 7, issue 14, 2012.
- Dan Flickinger et al. 2012: Dan Flickinger, Valia Kordoni, Yi Zhang, Antonio Branco, Kiril Simov, Petya Osenova, Catarina Carvalheiro, Francisco Costa, Sergio Castro. ParDeepBank: Multiple Parallel Deep Treebanking. In Iris Hendrickx, Sandra Kuebler and Kiril Simov (eds.) Proceedings of the 11th (TLT11), Lisbon, Portugal. Edicoes Colibri, pp. 97-108.

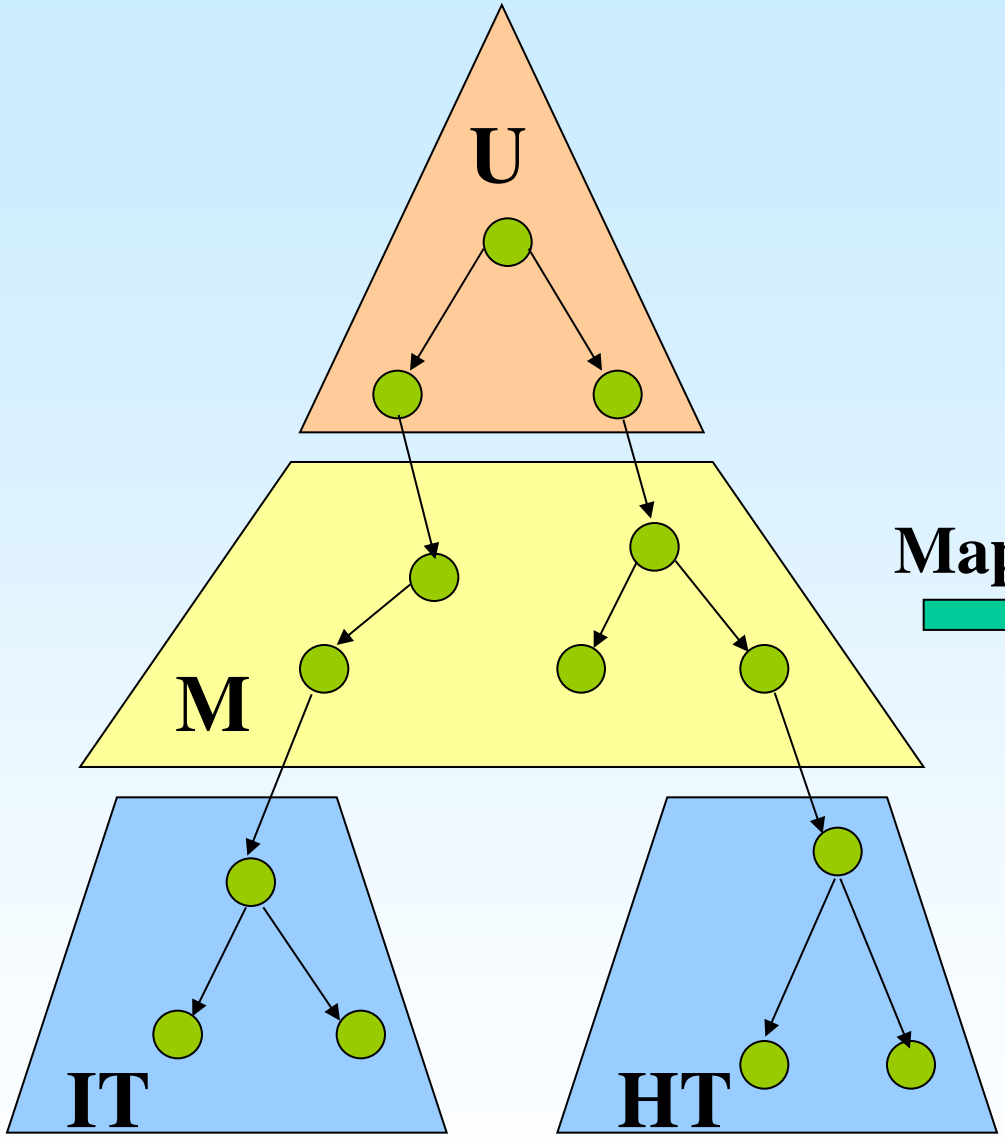
Other publications (2)

- Osenova 2009: Petya Osenova. On the Lexicographic Representation of Relational Nouns. In: Proceedings from the MONDILEX Workshop on Representing Semantics in Digital Lexicography, 29 June-1 July, Warsaw, Poland, pp. 205-210
- Simov and Osenova 2008: Kiril Simov and Petya Osenova. A Treatment of Coordination in the Bulgarian HPSG-based Treebank. In: *Proc. of FDSL 5, Peter Lang*, pp. 68-77.
- Simov and Osenova 2011: Kiril Simov and Petya Osenova. Towards Minimal Recursion Semantics over Bulgarian Dependency Parsing. In *Proceedings of Recent Advances in Natural Language Processing*, pages 471–478. Hissar, Bulgaria, 12-14 September 2011.
- Osenova 2014: Petya Osenova. Classifier noun phrases of the type N1N2 in Bulgarian. In: Olga Spevak (editor) Noun Valency, pp.141-160. John Benjamins Publishing Company.

Ontology-to-Text Relation Approach

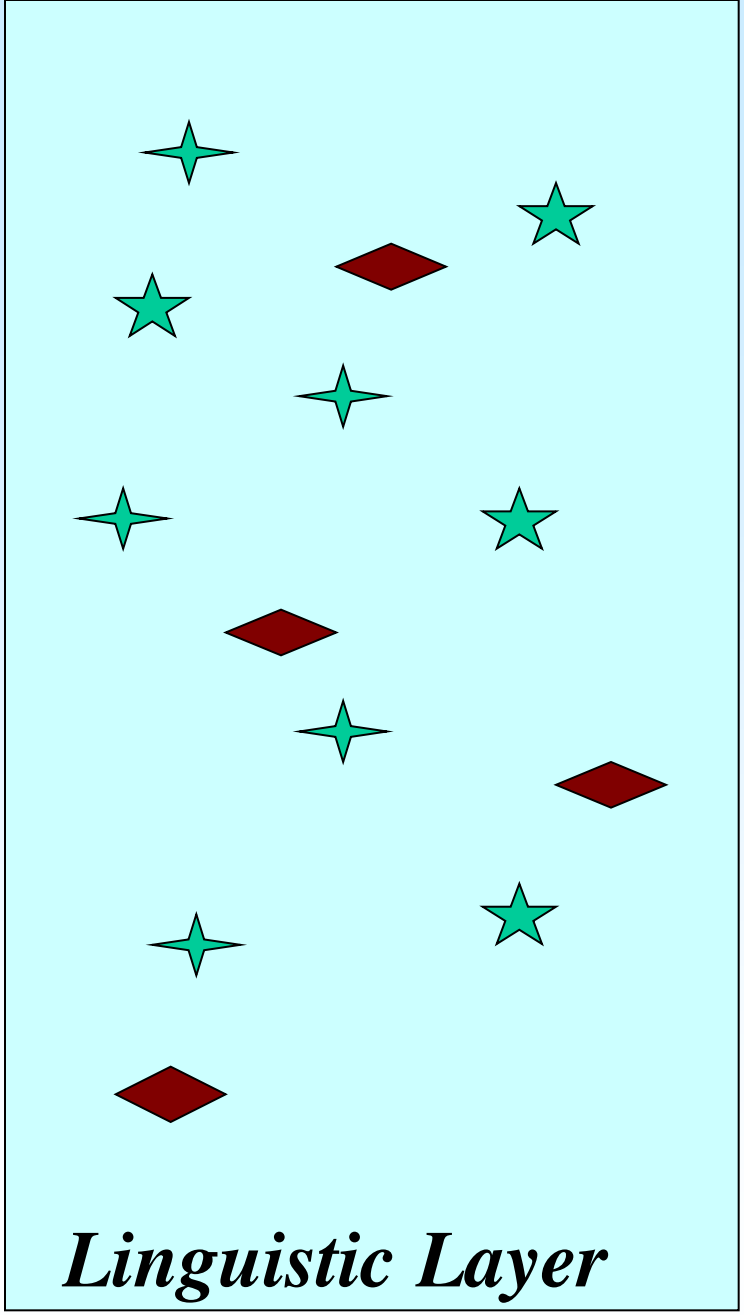
Based on Kiril Simov and Petya Osenova. 2008. Language Resources and Tools for Ontology-Based Semantic Annotation. In Proceedings of OntoLex 2008 Workshop at LREC 2008, pp. 9-13.

Work done in projects: AsIsKnown, LT4eL, LTfLL



Ontology

Mapping



Linguistic Layer

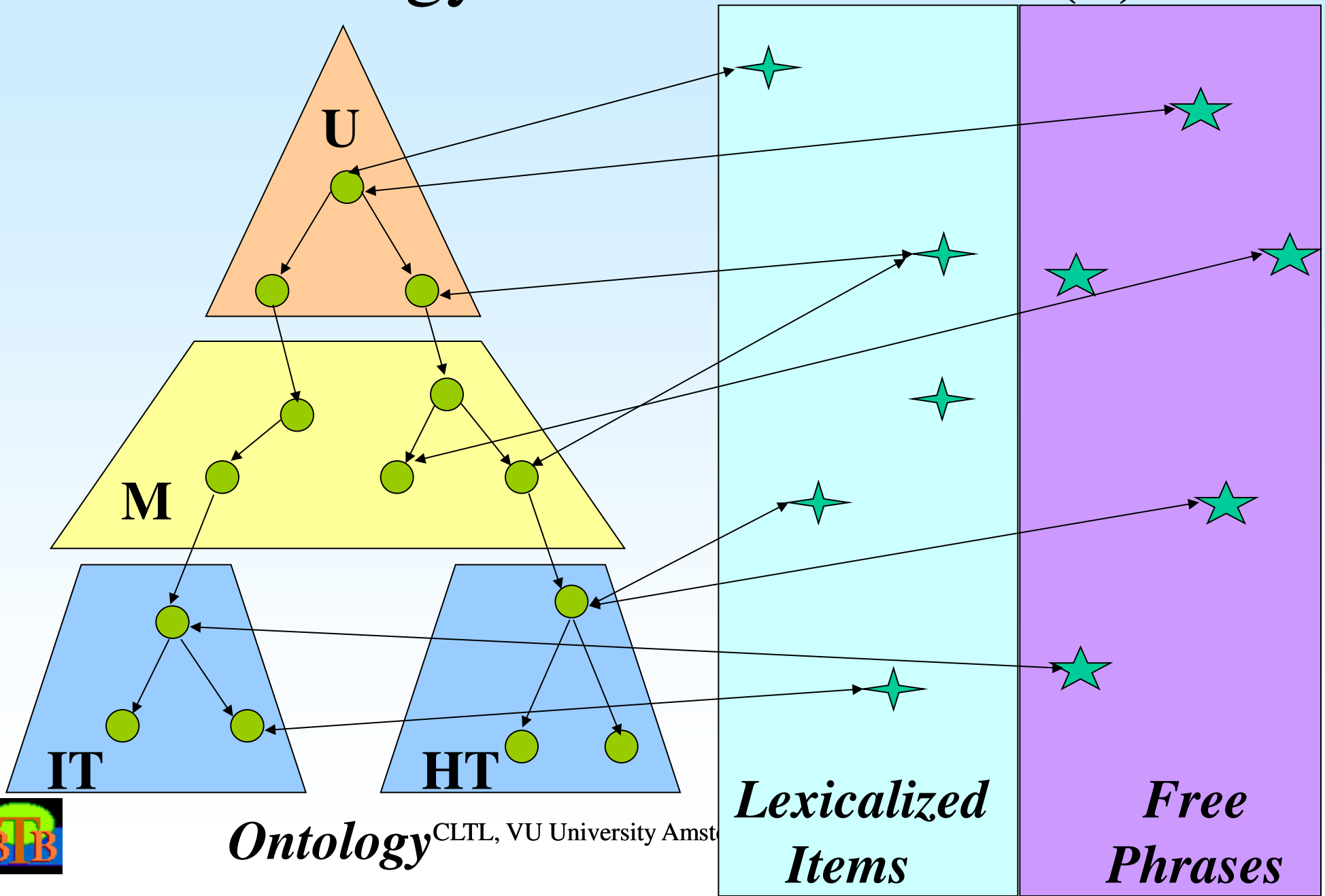
Lexicons and Annotation Grammars

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

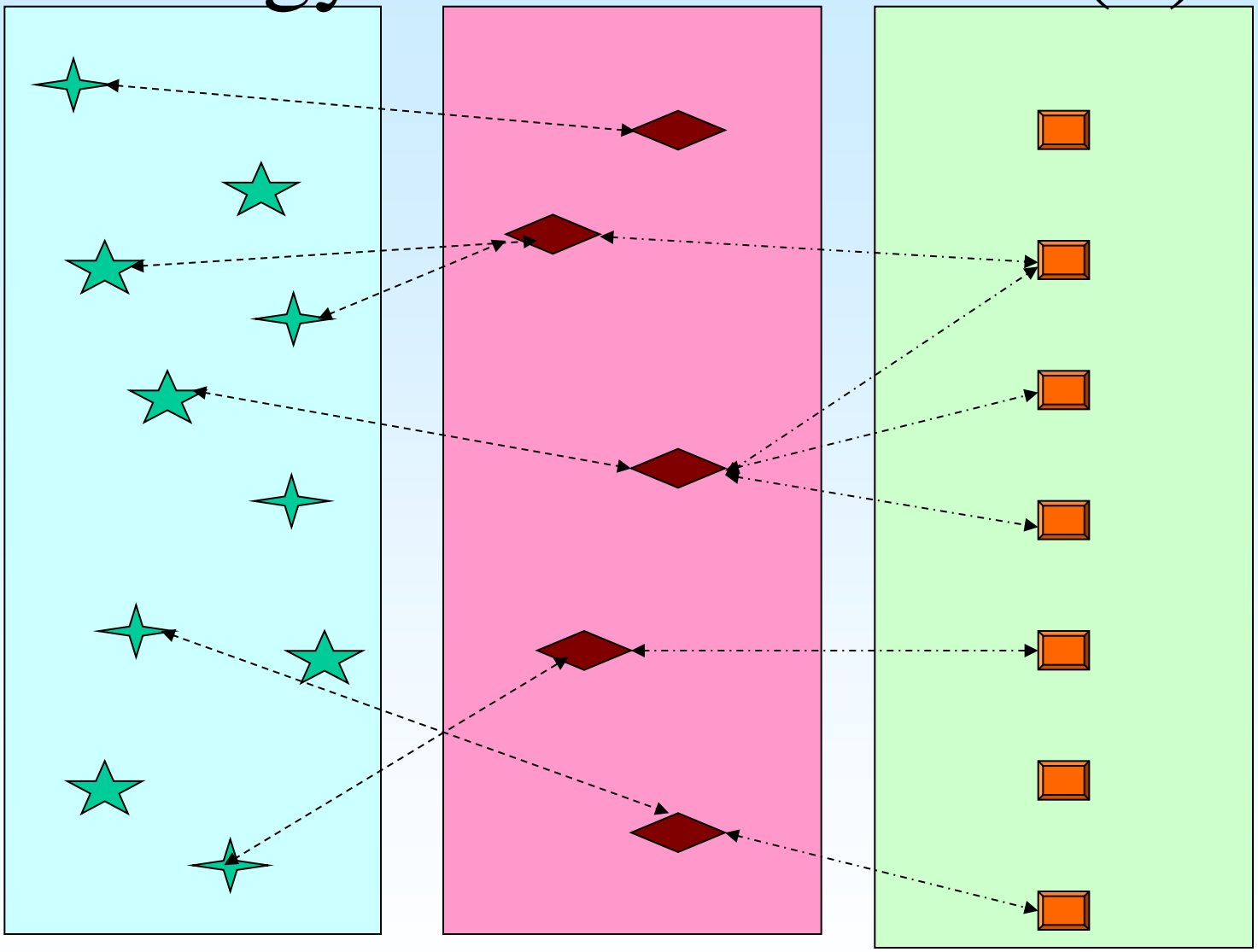
Ontology-to-Text Relation (1)

- 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)



Lexicon ***Grammar*** ***Text***

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

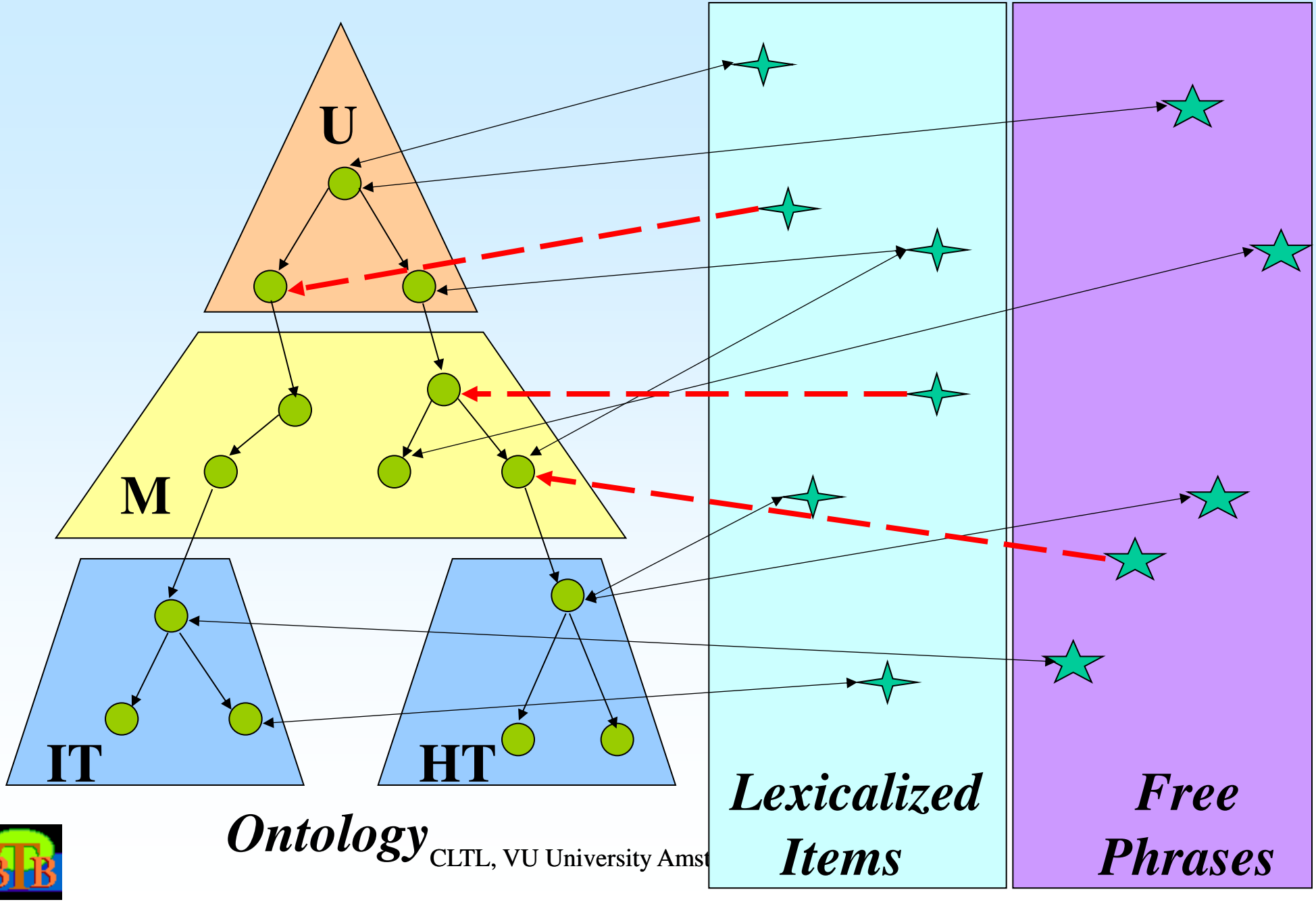
- Ideally, it is an extension of a deep grammar
- Minimally, it is a chunk grammar equipped with disambiguation rules for ambiguous terms
- The rules in the chunk grammar are created on the basis of the terms in the lexicon and rules from general chunk grammar
- Disambiguation rules are based on the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

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**



Ontology

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.

Bulgarian Ontology-based Lexicon

- The valence lexicon is a part the Bulgarian Ontology-based Lexicon (BOL) – (Simov and Osenova, 2010).
- The current version of BOL is based on DOLCE ontology (Masolo et a., 2003) extended with concepts from OntoWordNet (Gangemi et al., 2003) - a version 1.6 of WordNet aligned to DOLCE
- Intersection of EuroWordNet Base Concepts and Core WordNet (**1504 synsets**)
- Extended with lexical units extracted from the Bulgarian National Referent Corpus (**www.webclark.org**).

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

Other publications

- Osenova et. al 2008: Petya Osenova, Kiril Simov, Eelco Mossel 2008: Language Resources for Semantic Document Annotation and Crosslingual Retrieval. In: *Proc. of LREC 2008, ELRA*.
- Simov and Osenova 2010: Kiril Simov and Petya Osenova. Constructing of an Ontology based Lexicon for Bulgarian. In: Proceedings from LREC 2010, Malta, pp. 3840-3844.
- Osenova and Simov 2011: Petya Osenova and Kiril Simov. Syntactic-Semantic Treebank for Domain Ontology Creation. In: *Cognitive Studies*, vol. 11, Warsaw, pp. 213-225.
- Staykova et. al 2011: Kamenka Staykova, Gennady Agre, Kiril Simov and Petya Osenova. Language Technology Support for Semantic Annotation of Iconographic Descriptions. In: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 16.09.2011, at RANLP 2011, ISBN 978-954-452-019-9, pp. 51-56.

Bulgarian Language Technology

Based on

Kiril Simov, Petya Osenova, Sia Kolkovska,

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

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova and Kiril Simov. 2012. *Linguistic Analysis Processing Line for Bulgarian.* LREC 2012

Work started in projects: BulTreeBank and CLaRK (but continues ever after ☺)

BulTreeBank Language Technology

- **Tokenizers** - segmentation and classification
- **Morphological analyzers**
- **Disambiguator(s) - Lemmatizer**
- **Partial grammars**
  - Sentence splitter
  - Named-entity recognition module
  - Chunkers
- **Syntactic Parsers (current pipeline with POS tagger)**
- **Semantic Parser**
  - Partial RMRS annotator
  - (to be trained on Semantically Annotated BulTreeBank);

Morphological Dictionary Analyzer

- Assigns all possible analyses to the tokens
- Implemented as a regular grammar in CLaRK
- Works together with the ‘token classification’ and with the gazetteers
- Operates with 680 tags from the BulTreeBank tagset

(http://www.bultreebank.org/TechRep/BTB-TR03.pdf)

CLaRK System - [Root] Grammar - Slovnik

File Edit View DTD Definitions Tools Document Options Trees Help 261 Mb

[Root] Grammar - Slovnik - [DTD : grammars.dtd]

```
grammar
  name : Slovnik
  l:<tok aa="A-pd">\w<l c="A-pd" l="" w="ите"/></tok>:"небесносините"|"небивалите"|"неблаговъзпитаните"|"неблагонамере
  l:<tok aa="A-pd">\w<l c="A-pd" l="" w="те"/></tok>:"неандерталските"|"неаполитанските"|"неаполските"|"небоевите"|"не
  l:<tok aa="A-pd">\w<l c="A-pd" l="ен" w="йните"/></tok>:"неблагопристойните"|"недейните"|"недостойните"|"недълготрай
  l:<tok aa="A-pd">\w<l c="A-pd" l="ен" w="ните"/></tok>:"неактуалните"|"неакуратните"|"неапетитните"|"небесните"|"неб
    RE:"неактуалните"|"неакуратните"|"неапетитните"|"небесните"|"неблаговидните"|"неблагодарните"|"неблагозвучните"|"н
      "неактуалните"|"неакуратните"|"неапетитните"|"небесните"|"неблаговидните"|"неблагодарните"|"неблагозвучните"|"не
    RM:<tok aa="A-pd">\w<l c="A-pd" l="ен" w="ните"/></tok>
      <tok aa="A-pd">\w<l c="A-pd" l="ен" w="ните"/></tok>
  l:<tok aa="A-pd">\w<l c="A-pd" l="ин" w="ните"/></tok>:"неедните"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ъв" w="вите"/></tok>:"нееднаквите"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ък" w="ките"/></tok>:"негладките"|"неловките"|"немалките"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ър" w="рите"/></tok>:"недобрите"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ядък" w="едките"/></tok>:"нередките"
  l:<tok aa="A-pd">\w<l c="A-pd" l="як" w="еките"/></tok>:"непреките"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ял" w="елите"/></tok>:"недоузрелите"
  l:<tok aa="A-pd">\w<l c="A-pd" l="ям" w="емите"/></tok>:"неголемите"|"немите"
  l:<tok aa="A-pd;Ncmpd">\w<l c="A-pd" l="" w="ите"/><l c="Ncmpd" l="а" w="ите"/></tok>:"невежите"
  l:<tok aa="A-pi">\w<l c="A-pi" l="" w="и"/></tok>:"небесносини"|"небивали"|"неблаговъзпитани"|"неблагонамерени"|"неб
```

Attribute	Value
direction	right to left

/grammar[1]

thediction.com - SA Dictionary 2012

Start EN 18:45 8.7.2014 г.

MorphoSyntactic Tagger

- For each word form assign the right tag:
  Български (plural) книги <-> Български (singular) юнак
- The tagging is done over a large tagset – 680 tags (Each tag includes a part of speech information and morphosyntactic features – Ncmsi, …)
- A hybrid approach:
  – Guesser (currently a SVM Tagger)
  – Lexicon look-up and manual rules application
  – Statistical tagger (G-Tagger)

SVM Statistical POS Tagging

- SVMTool
  - Support Vector Machine (SVM) based tagger developed by Jesús Giménez and Lluís Márquez
  - Based on Vapnik's SVM implementaition *SVM Light* by Thorsten Joachims
  - Trained on a tagged corpus part of the BulTreeBank corpus (296K tokens) using additional dictionary (1 million word forms)
- Results
  - 93.5470% accuracy tested on a 12K homogenous corpus
  - 90.4710% accuracy tested on a 57K fiction corpus

Lexicon Look-up and Manual Rules

- Lexicon – Morphological Lexicon of Bulgarian (Popov, Simov, Vidinska, Osenova 2003) – 110000 lemmas, name list – 40000 names, abbreviations, parenthetical expressions
- The information from lexicons overwrites the suggestions of SVM tagger
- Manual rules with near to 100% accuracy (70)
  - *If ambiguous between a masculine count noun (Ncmt) and a singular short definite masculine noun (Ncmsh), Ncmt is chosen if the previous token is a numeral or a number.*

Example

Word	Tags
Той	**Ppe-os3m**
обаче	*Cc*; **Dd**
няма	Afsi; Vnitf-o3s; Vnitf-r3s; Vpitf-o2s; Vpitf-o3s; **Vpitf-r3s**
възможност	**Ncfsi**
да	*Ta*;**Tx**
следи	*Ncfpi*; *Vpitf-o2s*; *Vpitf-o3s*; **Vpitf-r3s**; *Vpitz–2s*
...	...

Table 1: **Sample fragment showing the possible tags suggested by the lexicon.** The tags that are further filtered by the rules are in italic; the correct tag is bold.

G-Tagger

- Guided learning
- Rich feature set (Shen & al. (2007))
- In the common application each token is associated with all tags
- In our architecture the output of SVM and Rules
- Two modes of application – soft and hard modes
- The best results for hard mode

Experiments

#	Lexicon (*source of*)	Linguistic Rules (applied to filter): (a) *the lexicon features*	(b) *the output tags*	Beam size	Accuracy (%) Sentence-level	Token-level
1	–	–	–	1	52.95	95.72
2	–	–	yes	1	64.50	97.20
3	features	–	–	1	70.40	97.83
4	features	yes	–	1	70.30	97.80
5	features	yes, *for test only*	–	1	70.40	97.84
6	features	–	yes	1	71.34	97.91
7	features	yes	yes	1	71.69	97.93
8	features	yes	yes	3	71.94	97.98

Table 4: **Evaluation results on the test dataset.** Line 1 shows the evaluation results when using features derived from the text corpus only; these features are used by all systems in the table. Line 2 further uses the contextual linguistic rules to limit the set of possible POS tags that can be predicted. Note that these rules (1) consult the lexicon, and (2) always predict a single POS tag. Line 3 uses the POS tags listed in the lexicon as features, i.e., as soft suggestions only. Line 4 is like line 3, but the list of feature-tags proposed by the lexicon is filtered by the contextual linguistic rules. Line 5 is like line 4, but the linguistic rules filtering is only applied at test time; it is not done on training. Lines 6 and 7 are similar to lines 3 and 4, respectively, but here the linguistic rules are further applied to limit the set of possible POS tags that can be predicted, i.e., the rules are used as hard constraints. Finally, line 8 is like line 7, but here the beam size is increased to 3.

Comparison to Other Works

Tool/Authors	Method	# Tags	Accuracy (token-level, %)
*TreeTagger	Decision Trees	680	89.21
*ACOPOST	Memory-based Learning	680	89.91
*SVMtool	Support Vector Machines	680	92.22
*TnT	Hidden Markov Model	680	92.53
(Georgiev et al., 2009)	Maximum Entropy	680	90.34
(Simov and Osenova, 2001)	Recurrent Neural Network	160	92.87
(Georgiev et al., 2009)	Maximum Entropy	95	94.43
(Savkov et al., 2011)	SVM + Lexicon + Rules	680	94.65
(Tanev and Mitkov, 2002)	Manual Rules	303	95.00(=P=R)
(Simov and Osenova, 2001)	Recurrent Neural Network	15	95.17
(Dojchinova and Mihov, 2004)	Transformation-based Learning	40	95.50
(Dojchinova and Mihov, 2004)	Manual Rules + Lexicon	40	98.40
	Guided Learning	680	95.72
	Guided Learning + Lexicon	680	97.83
This work	Guided Learning + Lexicon + Rules	680	97.98
	Guided Learning + Lexicon + Rules	*49*	*98.85*
	Guided Learning + Lexicon + Rules	*13*	*99.30*

Table 5: **Comparison to previous work for Bulgarian.** The first four lines report evaluation results for various standard POS tagging tools, which were retrained and evaluated on the BulTreeBank. The following lines report token-level accuracy for previously published work, as compared to our own experiments using guided learning.

Typical Errors

- 711 errors
- 206 not problematic for lemmatization
- 190 not problematic for parser
- Many errors related to acronyms and names

Freq.	Gold Tag	Proposed Tag
43	Ansi	Dm
23	Vpitf-r3s	Vnitf-r3s
16	Npmsh	Npmsi
14	Vpiif-r3s	Vniif-r3s
13	Npfsd	Npfsi
12	Dm	Ansi
12	Vpitcam-smi	Vpitcao-smi
12	Vpptf-r3p	Vpitf-r3p
11	Vpptf-r3s	Vpptf-o3s
10	Mcmsi	Pfe-os-mi
10	Ppetas3n	Ppetas3m
10	Ppetds3f	Psot–3–f
9	Npnsi	Npnsd
9	Vpptf-o3s	Vpptf-r3s
8	Dm	A-pi
8	Ppxts	Ppxtd
7	Mcfsi	Pfe-os-fi
7	Npfsi	Npfsd
7	Ppetas3m	Ppetas3n
7	Vnitf-r3s	Vpitf-r3s
7	Vpitcam-p-i	Vpitcao-p-i

Table 6: **Most frequently confused pairs of tags.**

Joint Model for Ensemble of MorphoSyntactic Taggers and Dependency Parsers

Kiril Simov, Ginka Ivanova, Maria Mateva and Petya Osenova. 2013. *Integration of Dependency Parsers for Bulgarian*. TLT12;

Kiril Simov, Ginka Ivanova, Maria Mateva and Petya Osenova. 2014. *A System for Experiments with Dependency Parsers for Bulgarian*. LREC 2014;

Iliana Simova, Dimitar Vasilev, Alexander Popov, Kiril Simov, Petya Osenova. 2014. *Combining Morphosyntactic Taggers and Dependency Parsers for Bulgarian*. In SPMRL-SANCL WS at COLING 2014.

Motivation on Joint Model for Several Tasks: NLP side

(1) Avoiding the accumulation of errors inherent to pipeline processing,

(2) Overcoming the low speed of model-chaining approaches,

(3) Confirming the success of previous developments in joint modeling

Motivation on Joint Model for Several Tasks: Linguistic Side

- Assessing the benefits of modeling the interactions that exist among different linguistic levels. Examples from Bulgarian:
  - Long-distance agreement – subject-verb, verbal clitics-explicit object/indirect object
  - Unexpressed subjects participate in co-reference chains of control, binding, etc. constructions
  - Agreement between head noun and relative pronouns
  - Agreement between secondary predication and verb argument

Motivation on Ensemble Model for Parsing

McDonald and Nivre, 2007: They showed that transition-based and graph-based frameworks made different errors on the same dataset

- **Ensemble systems**: weighted combinations of both systems output
- **Hybrid systems**: design of a single system integrating the strengths of each framework
- **Novel approaches**: based on combination of new training and inference methods

Similarly for POS tagging (MS Tagging)

Parsing Combination

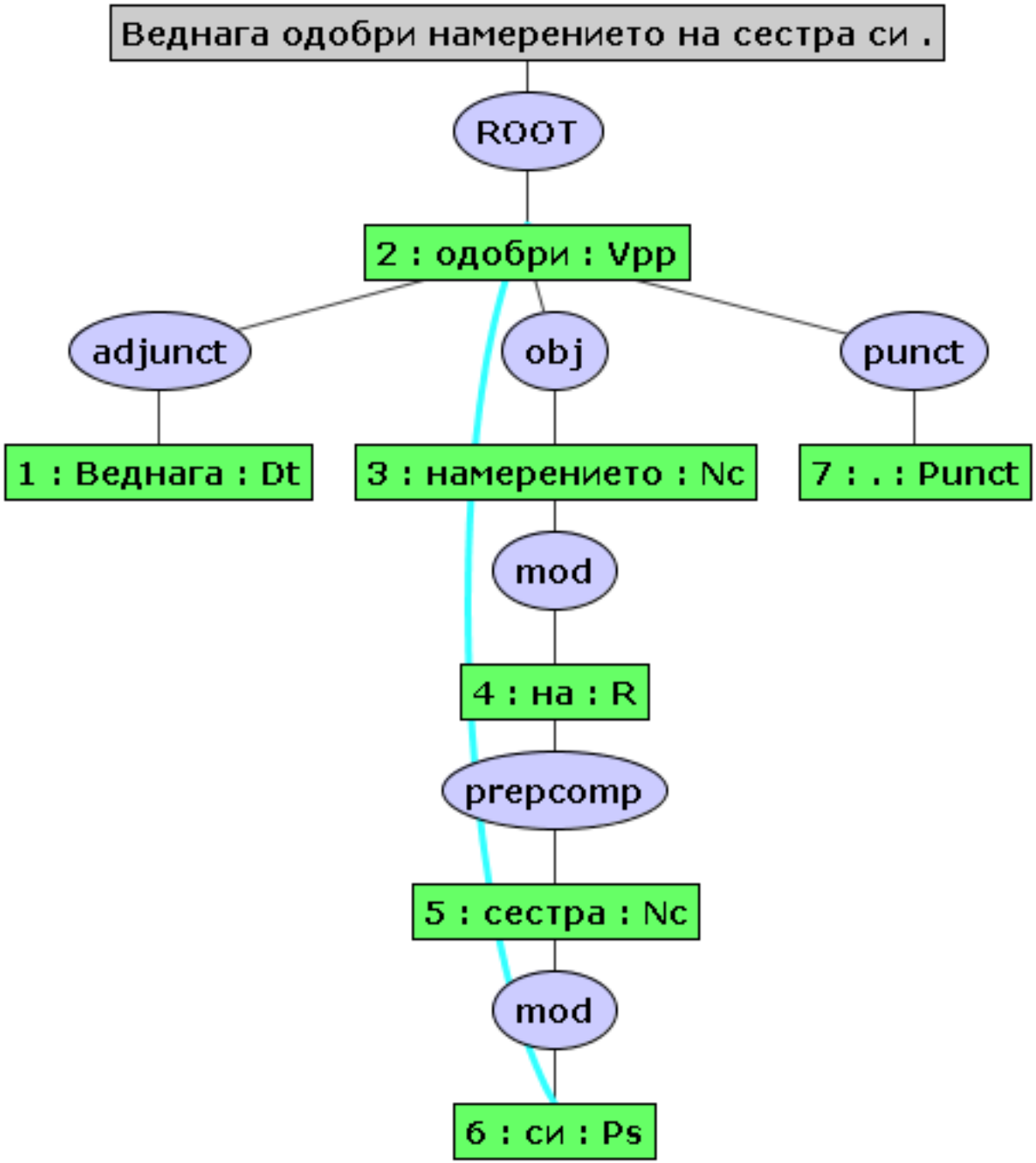
Surdeanu and Manning, 2010

- Different approaches to ensemble of parsing
- Voting – Selection of arcs on the basis of several models trained on the same datasets: majority voting, weighted voting
- Machine Learning (meta-classification) – learning which arc is better choice: majority and minority dependencies

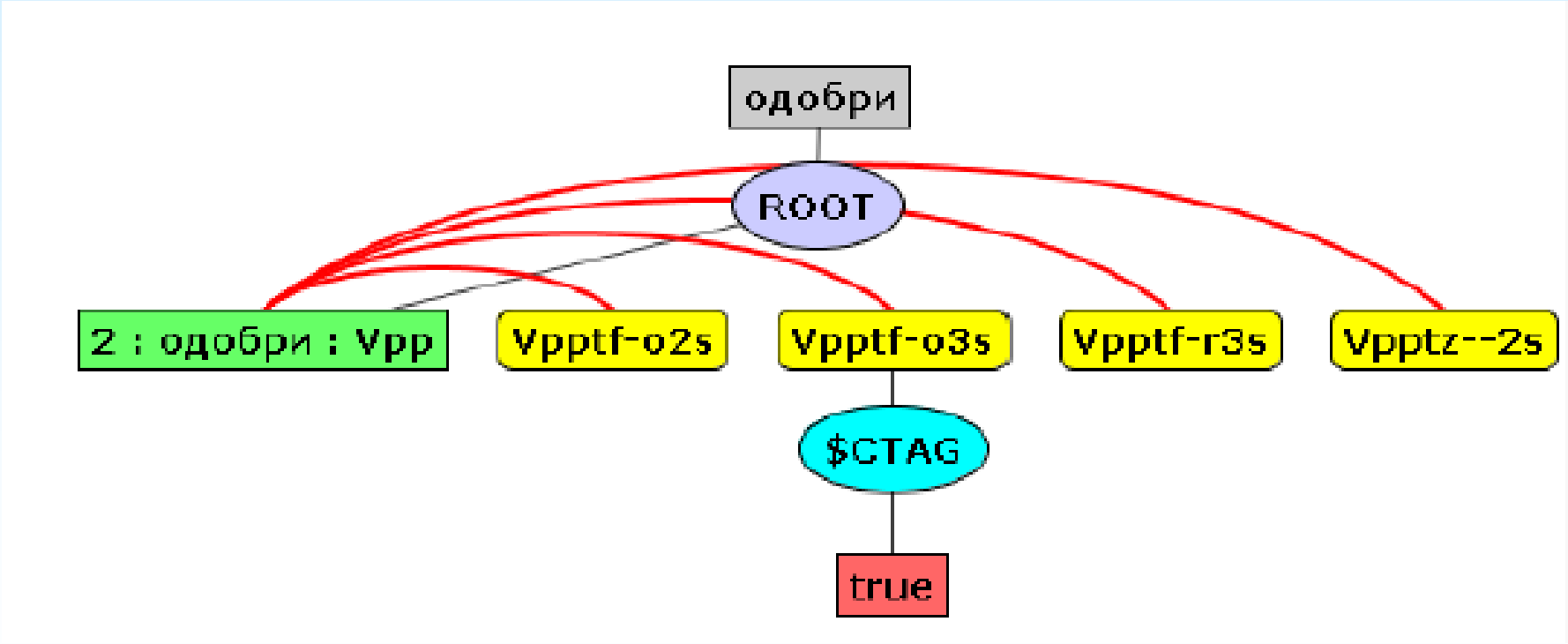
Tagging and Parsing Combination

- In order to model the interaction of morphological characteristics and syntactic structures we encode morphological characteristics as an extension of the dependency trees
- The features that are used for ranking the MS tags and dependency arcs are the same

Dependency Tree



Subtree of the candidate MS tags and the correct tag



Construction of Tree – LocTr

Attardi and Dellorletta, 2009:

- The construction of the tree starts with an empty tree containing only a root
- On each steps LocTr is looking for the arc with the best rank that can extend the current partial tree
- The result tree is only locally optimized

Construction of Tree – GloTr

The construction of the tree follows ideas of MST Parser for non-projective parsing represented in MacDonald 2006

- The algorithm starts with a complete dependency
- The arcs are ranked depending on the method
- Chu-Liu-Edmonds algorithm for maximal spanning tree select an optimal tree
- The optimization is global

Tagger/Parsing Models and ParseBanks

- The Treebank was divided in 10 parts
- Combining in 10 training and test datasets: Each training set contains 10790 sentences and each test set – 1198
- Each tagger is trained on each of the training sets and evaluated on the test set and the results are stored
- Each parsing model was trained on each of the training sets and evaluated on the corresponding test set with tags for each tagger and the result is stored in the corresponding parsebank

Taggers and Parsers

- Taggers: BLL (95.91%), Mate (94.92%), TreeTagger (93.12%)
- Parsers: MaltParser – two models, MateParser – one model, MSTParser – one model, TURBOParser – one model
- 15 combinations of Tagger+Parser models

Parser Performance – UAS

MLT07	MLT09	MATE01	MST05	Turbo02	training data
0.900	0.908	0.929	0.911	0.927	gold
0.881	0.890	0.910	0.890	0.911	BLL tagger
0.881	0.889	0.908	0.890	0.910	Mate tagger
0.857	0.865	0.883	0.865	0.883	TreeTagger

Voting Combination

	Models	Algorithm	Rank01 Number		Rank02 Sum		Rank03 Average	
			LAS	UAS	LAS	UAS	LAS	UAS
(1)	all models	LocTr	88.55	92.05	88.61	92.10	85.57	88.82
		GloTr	88.55	91.96	88.65	92.04	84.54	88.75
(2)	all Mate01 and Turbo02 models	LocTr	87.68	91.38	87.80	91.48	86.94	90.55
		GloTr	87.58	91.21	87.82	91.45	86.84	90.62
(3)	best combination	LocTr	88.90	92.34	89.05	**92.47**	86.19	89.40
		GloTr	88.94	92.31	89.14	**92.45**	85.23	89.27

Combining Parses by Machine Learning

- RandomForest in R:
  - it supports classification and regression methods; and
  - it does not overfit
- The Parsebanks were divided again into training and test parts (90/10)
- Our goal was to evaluate each tag/arc suggested by a tagging/parsing model as a correct or an incorrect tag/arc for a given context

Features

- Features for arcs: relation (**Rel**), distance to the parent node measured as a number of words in the sentence between the two nodes (**Dist**) and direction of the parent node (**Dir**)
- The features for each word include: the word form (**WF**), lemma (**Lemma**), morphosyntactic tag (**POS**)

Feature Vector

Feature	Value
Word	the current node
WordBefore	the word before the current node
WordAfter	the word after the current node
ParentWord	the parent word
PWordBefore	the word before the parent word
PWordAfter	the word after the parent word
SelectedArc	one of the arcs suggested by one of the models for the node
SelectedTag	one of the arcs suggested by one of the models for the node
CorrectIncorrect	true or false depending on whether the selected pair is the correct one for the node

Classification vs. Regression

- We first tried RandomForest classification – too coarse:

	Correct	Incorrect
True	65.2%	73.4%
False	34.8%	26.6%

- Regression uses a continuous interval of values – fine grained ranking

Results for Parsing

Model	Algorithm	LAS	UAS
all	LocTr	89.17	92.46
	GloTr	89.23	92.27
all Mate01 and Turbo02 models	LocTr	88.26	91.81
	GloTr	88.32	91.87
best combination	LocTr	89.76	93.18
	GloTr	89.81	93.22

Results for Tagging

Voting	Rank01 Number	Rank02 Sum	Rank03 Average
BLL, Mate, TreeTagger	96.24	96.24	95.22

MLearning	WMax	WMin	WSum
BLL, Mate, TreeTagger	96.10	96.20	96.25
BLL, Mate	96.62	96.59	**96.63**
BLL, TreeTagger	95.89	96.08	96.09
Mate, TreeTagger	95.29	95.40	96.25

System for Experiments with Taggers and Parsers

- The best result requires evaluation of all combinations. For 15 models - 32738. Depends on many parameters.
- This motivated us to start implementation of a system which provides facilities for the following tasks:
  - Treebank declaration (versions, division in test and train…)
  - Training of initial models (what features, versions,…)
  - Ranking of arcs
  - Combination of models
- Ideally representing all the data as LOD for easy reuse and sharing

Current Projects

- **EUCases**: EUropean and National CASE Law and Legislation Linked in Open Data Stack
  - Annotation of Case Protocols with Ontological information and converting to LOD
- **QTLeap**: Quality Translation by Deep Language Engineering Approaches
  - Exploiting Deep Linguistic Analisys for MT transfer
  - (R)MRS structure over dependency parses augmented with ontology and instance information from LOD

Thank you for your attention!