BulTreeBank Group: Updates

Petya Osenova

Sofia University “St. Kl. Ohridski”, IICT-BAS

29.07.2013

Acknowledgement

Petya Osenova's participation is supported by the FP7 Capacity Programme:

AComIn: Advanced Computing for Innovation, hosted at IICT-BAS

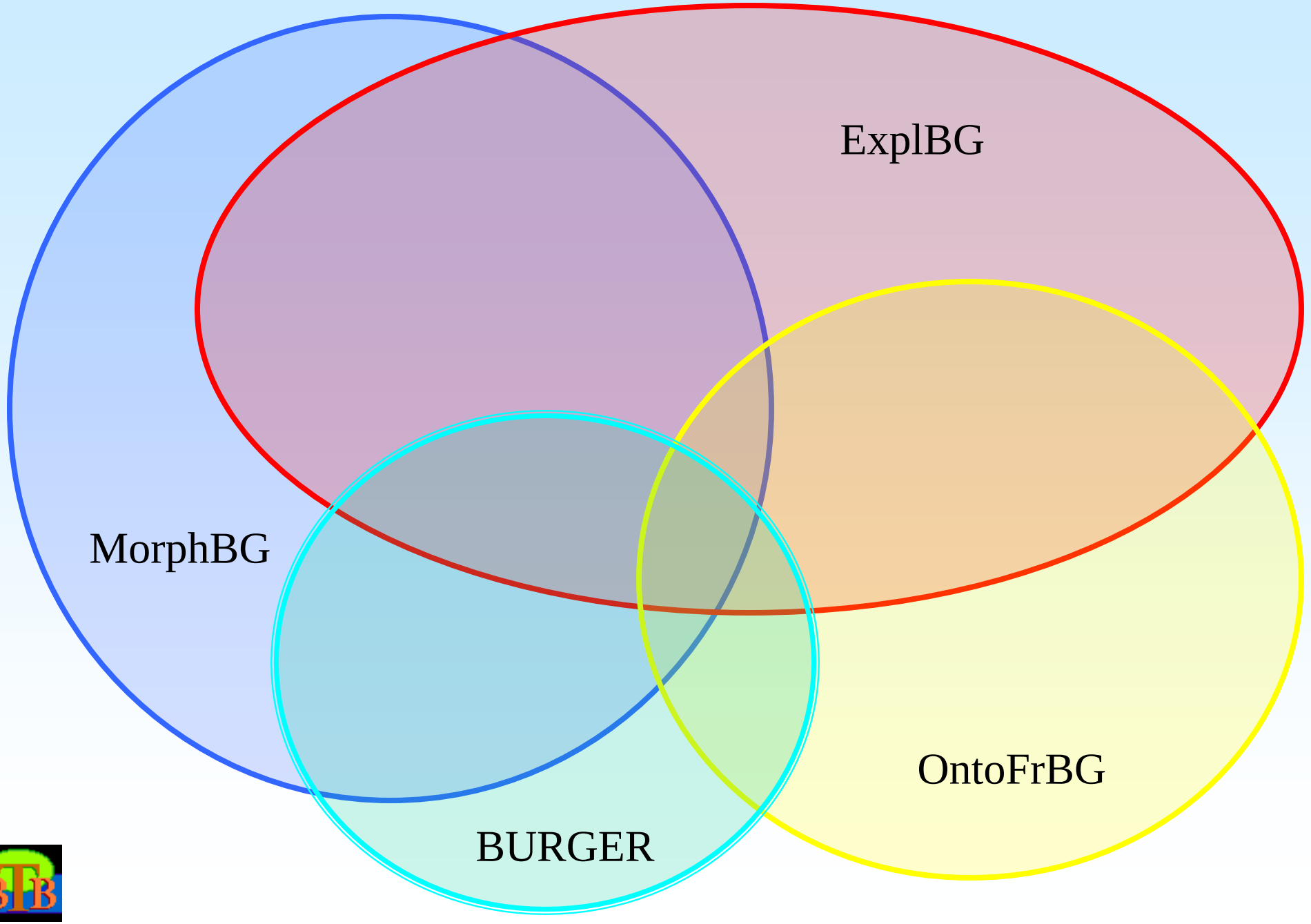
Grant Agreement: 316087

Plan of the Talk

- Lexicon database
- Dependency parsing
  - Combination of three tasks: 3in1
  - Combination of several parsers
- New projects

Joint Database for Several Lexicons

- The goal is to have a joint schema for the following lexicons:
  - Morphological lexicon
  - Ontology-based and Valency lexicon
  - Explanatory dictionary of Bulgarian
  - BURGER lexicon
- Each sense is connected with the right conceptual information, morphological paradigm, valency frames, HPSG types


ExplBG
MorphBG
BURGER
OntoFrBG

Extraction of BURGER Lexicon

- The mapping between LKB types to other information in the lexicon is semi-automatic
- The main problems are:
  - homonymy,
  - granularity of the description of some phenomena,
  - interaction with the grammar
- Other applications: lexicon for dependency parsing, lexicon for semantic annotation

Interaction with the Grammar

- The lexical entries are connected with elements of the grammar – lexical types, paradigm types and irules
- When extracting lexicon for the grammar – the program extracts the minimal part of the paradigm types and corresponding rules. In this way only the necessary linguistic knowledge is loaded into the grammar

3in1: Combining POS tagging, Dependency Parsing and Coreference Resolution

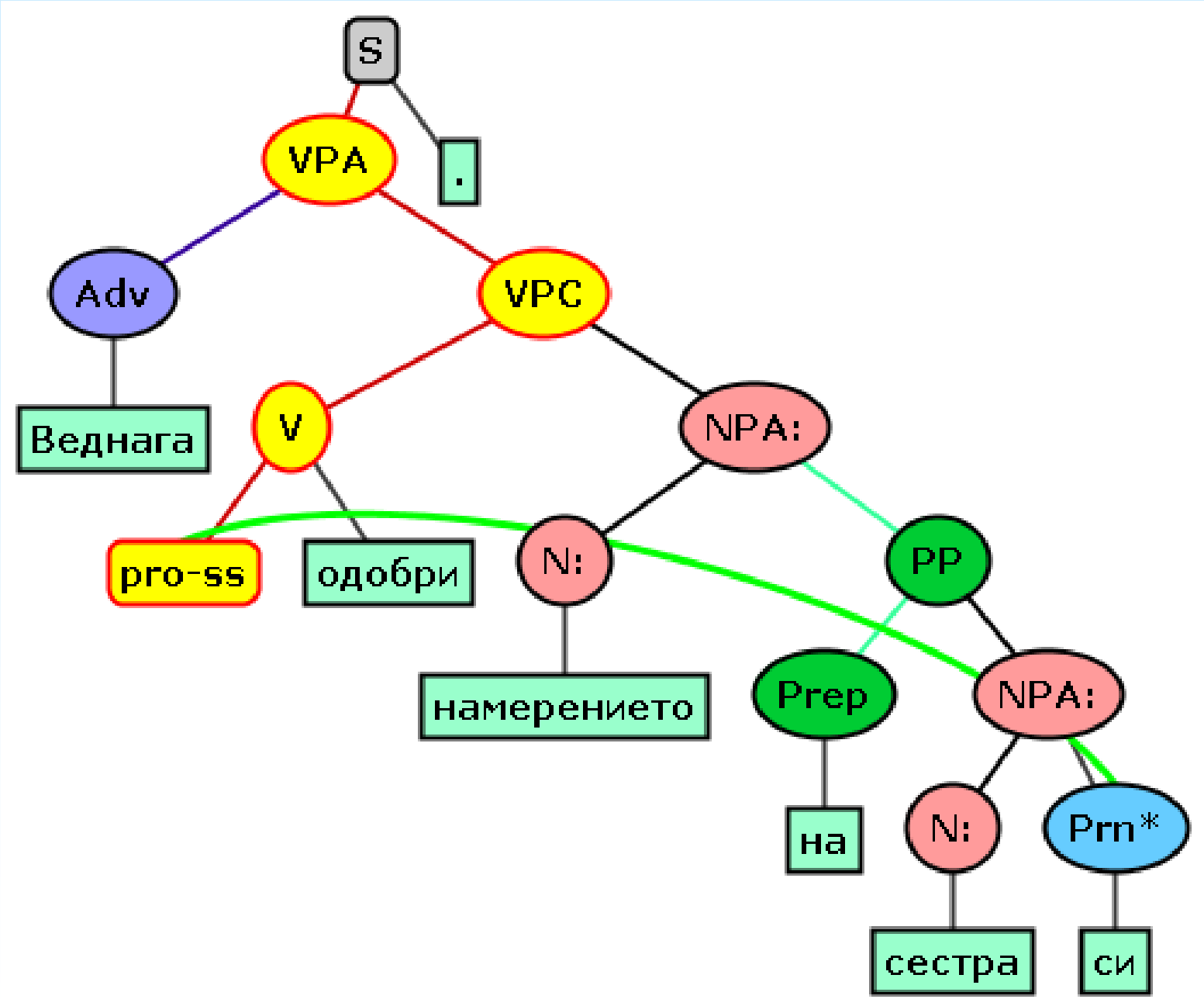
- This is a paper, accepted at RANLP 2013
- Data:
  - annotated sentences from BulTreeBank – converted to a dependency format.
  - inflectional lexicon of Bulgarian
  - morphological guesser, which narrows down the candidate POS tags for each word to manageable numbers.

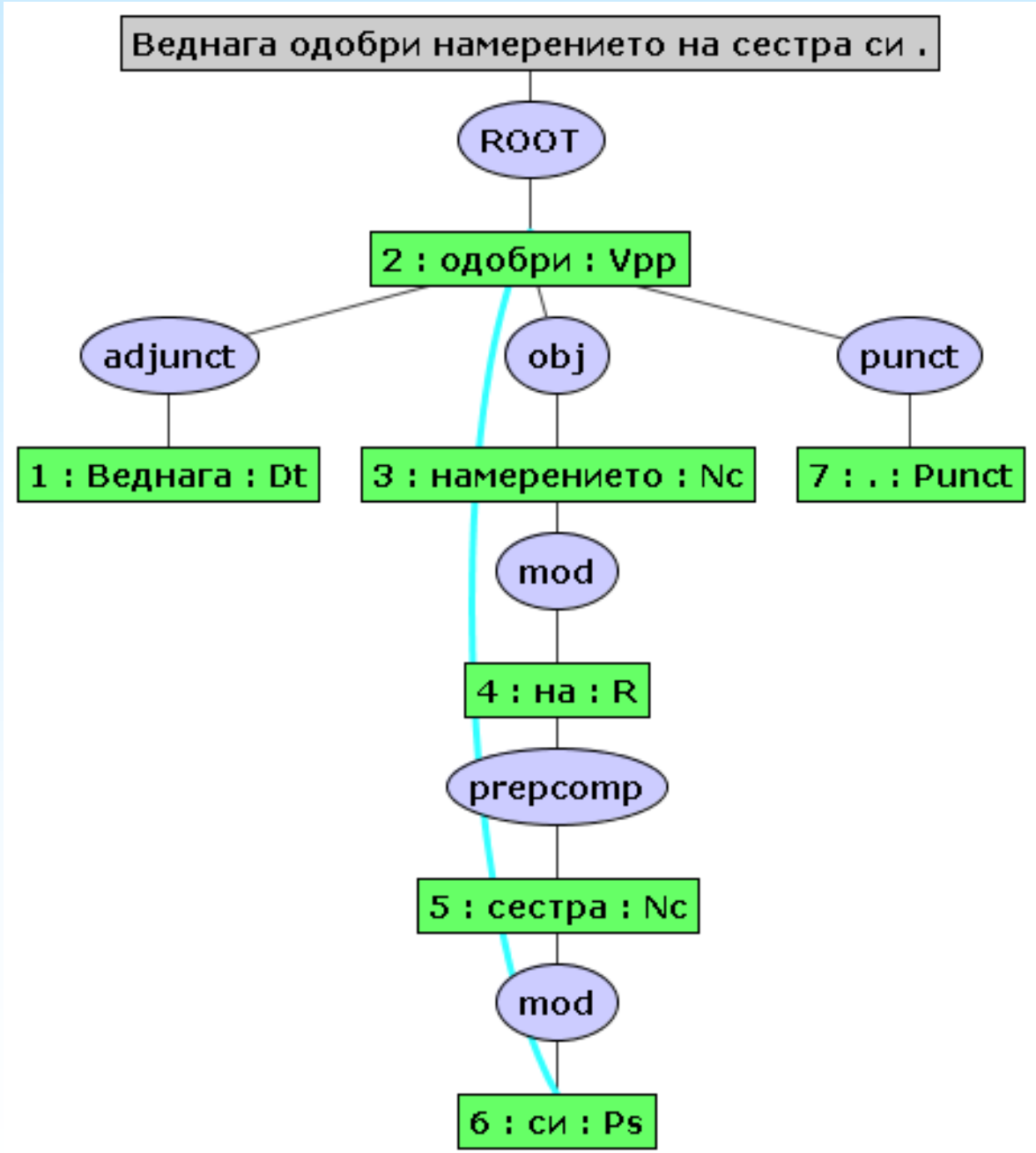
Why Combining the Tasks?

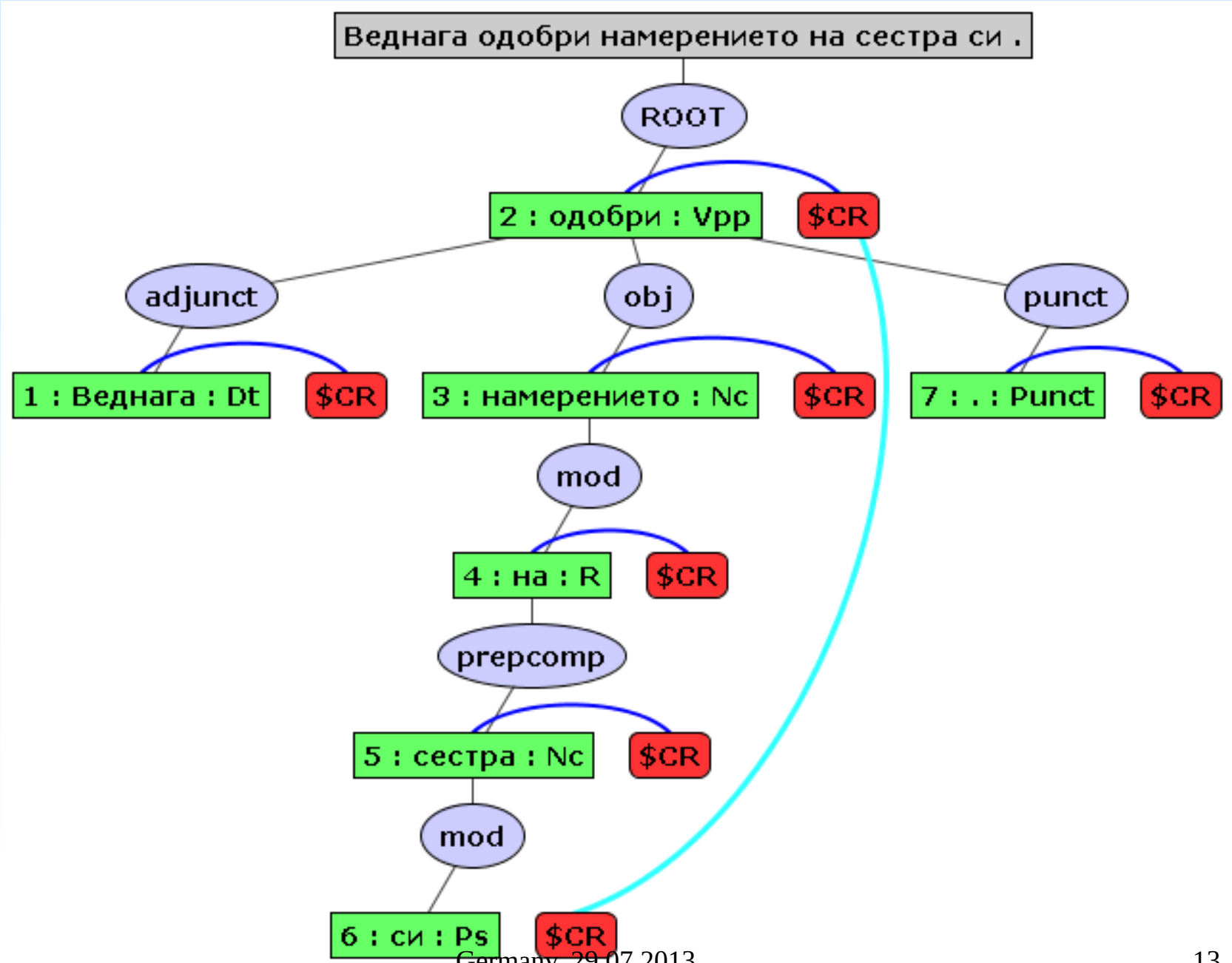
- Avoiding the accumulation of errors inherent to pipeline-based processing,
- Overcoming the low speed of model-chaining approaches,
- Confirming the success of previous developments in joint modeling against a new language dataset;
- Assessing the benefits of modeling the interactions that exist among morphology, syntax and discourse.

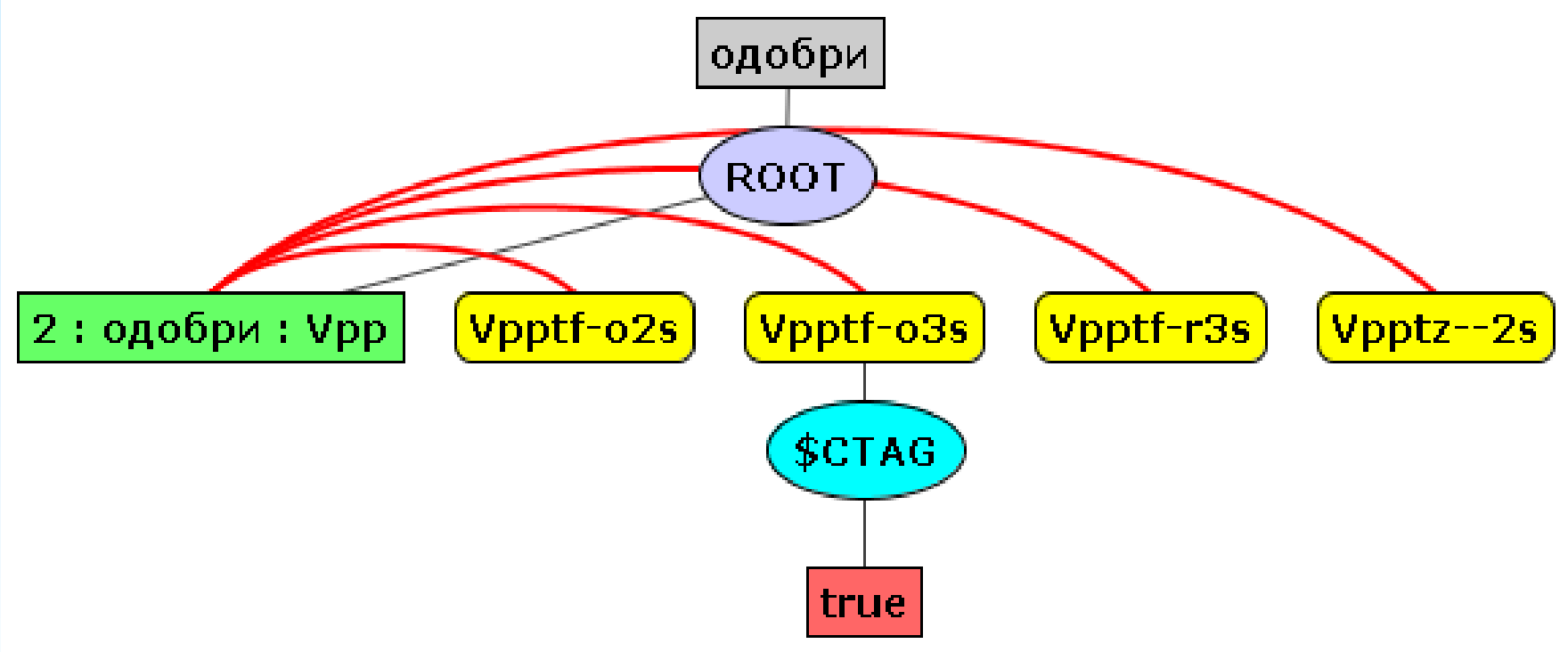
Strategy

We create an extended dependency tree that incorporates service nodes and links, through which additional knowledge, such as POS tag candidates, correct POS tags and co-reference relations, can be fed into the MSTParser algorithm for nonprojective dependency parsing (McDonald et al., 2005).


S
VPA
.
Adv
VPC
Веднага
V
NPA:
pro-ss
одобри
N:
PP
намерението
Prep
NPA:
на
N:
Prn*
сестра
си


Веднага одобри намерението на сестра си .
ROOT
2 : одобри : Vpp
adjunct
obj
punct
1 : Веднага : Dt
3 : намерението : Nc
7 : . : Punct
mod
4 : на : R
prepcomp
5 : сестра : Nc
mod
6 : си : Ps


Веднага одобри намерението на сестра си .
ROOT
2 : одобри : Vpp
$CR
adjunct
obj
punct
1 : Веднага : Dt
$CR
3 : намерението : Nc
$CR
7 : . : Punct
$CR
mod
4 : на : R
$CR
prepcomp
5 : сестра : Nc
$CR
mod
6 : си : Ps
$CR


одобри
ROOT
2 : одобри : Vpp
Vpptf-o2s
Vpptf-o3s
Vpptf-r3s
Vpptz--2s
$CTAG
true

Maximum Spanning Tree Model

- A complete graph is constructed: words as nodes, arcs between each pair of nodes with all possible labels. Each arc has a weight predicted by a learner
- A spanning tree with a maximal weight is selected
- We have added a filter mechanism which deletes the inappropriate arcs

Examples of Filtering Rules

- We assign empty feature vectors to the dependency arcs that do not comply with either of the following preconditions:
  - root nodes can only be linked to word nodes;
  - word nodes can only be linked to their corresponding co-reference ($CR) and POS candidate ($TAG) nodes, other word nodes

Incorporated Features

- Word-Sentence Root; Word-Word; Co-reference-Word……
- Example:

Co-reference-Word: string; POS-tag candidates for the corresponding word form

Results

- Dataset comprises 190 000 tokens
- Of these, we used 90% for training, and 10% – for testing.
- We compiled the two subsets by allocating every tenth sentence to the test split, and putting all remaining sentences into the training split.

Results (2)

#	System	POS	Co-reference			Dependency		
		Accuracy (%)	Prec (%)	Recall (%)	F	LAS (%)	UAS (%)	LA (%)
1	features&morph	95.99	80.90	33.08	46.96	81.22	85.12	88.96
2	features&decomp. morph*	95.52	81.04	32.08	45.96	80.50	84.55	88.59
3	1&word context	95.95	80.97	33.23	47.12	81.42	85.35	88.95
4	3&distances	95.98	82.03	37.06	**51.05**	81.82	85.70	89.32
5	4&context-bigrams	97.12	81.77	35.38	49.39	82.29	86.19	89.65
6	5&additional conjunctions	**97.13**	81.16	34.30	48.22	**82.39**	86.17	89.64

Combination of Dependency Parsers

- Combination of results from several parsers (Nivre and McDonald 2008)
- Creation of corpus with results from 14 different models of Malt and MST parsers
- Voting strategies
- Machine learning over the created corpus
- The best result: 92.45 % UAC and 89,56 % LAC
- 1,97 % wrong arcs for all parsers

Combination of Dependency Parsers (2)

- Two approaches to tree construction from several parses:
  - MST approach (global optimization) – filtering on the basis of arc weight
  - Linear Tree Combination (local optimization) – (Attardi and Dell'Orletta 2009)
- Voting – number of parsers that produced a given arc; the accuracy of the parsers – average and sum
- Machine Learning – predicts the weight of each arc

New Projects

- **QTLeap** (2013-2016) – Quality Translation by Deep Language Engineering Approaches
- **EUCases** (2013-2015) – EUropean and National Legislation and CASE Law Linked in Open Data Stack
- **ProMoRe** (2014-2017) – Process Modeling Repository

EUCases

- Partners: empirica (DE), APIS (BG), IICT-BAS (BG), University of Torino (IT) Averbis (DE), Nomotika (IT)
- Transforming multilingual legal open data into linked open data after semantic and structural analysis
- Our involvement: Ontology construction, Multilingual Semantic Annotation, Multilingual Semantic Search and NLP tools for Bulgarian

ProMoRe

- Partners: Fluid Operations (DE), University of Applied Sciences Mannheim (DE), APIS (BG), IICT-BAS (BG)
- Ontology-based service repository in the cloud, populated with web services
- Our involvement: Ontology creation, definition of workflow for NLP services, NLP services for Bulgarian