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)


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/
<table>
<thead>
<tr>
<th>name</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>TüBa-D/DC: <em>Der Schatz im Silbersee</em>, Karl May</td>
<td>Constituency</td>
</tr>
<tr>
<td>Automatically parsed edition of &quot;Der Schatz im Silbersee&quot; (1894) by Karl May, from the TüBa-D/DC.</td>
<td></td>
</tr>
<tr>
<td>HPSG-based Syntactic Treebank of Bulgarian</td>
<td>Dependency</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>
#4713: Трябва ли президентът да бъде избран от парламента?
Welcome to INESS

INESS is an open system serving a range of research needs, offering an interactive, language independent platform for building, accessing, 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 LFG. Part of the corpus will be efficiently manually disambiguated with the LFG Parsebanker, while the rest will be stochastically 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 the 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 your IDP (please use the eduGAIN link). Otherwise, you will have to register an OpenID account. After you have logged in at INESS for the first 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.
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 · Ifg (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
<table>
<thead>
<tr>
<th>Id</th>
<th>Uid</th>
<th>Words</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1844</td>
<td>2139060</td>
<td>9</td>
<td>Народното събрание осъществява законодателната власт и упражнява парламентарен контрол.</td>
</tr>
<tr>
<td>1845</td>
<td>2139061</td>
<td>8</td>
<td>Народното събрание се състои от 240 народни представители.</td>
</tr>
<tr>
<td>1846</td>
<td>2139062</td>
<td>9</td>
<td>Народното събрание се избира за срок от четири години.</td>
</tr>
<tr>
<td>1847</td>
<td>2139063</td>
<td>29</td>
<td>За народен представител може да бъде избран български гражданин, който няма друго гражданство, навършил е 21 години, не е поставен под запрещение и не изпълнява наказание лишаване от свобода.</td>
</tr>
<tr>
<td>1848</td>
<td>2139064</td>
<td>14</td>
<td>Кандидатите за народни представители, които заемат държавна служба, прекъсват изпълнението й след регистрацията си.</td>
</tr>
<tr>
<td>1849</td>
<td>2139065</td>
<td>15</td>
<td>Законността на изборите може да се оспори пред Конституционния съд по реда, определен със закон.</td>
</tr>
<tr>
<td>1850</td>
<td>2139066</td>
<td>6</td>
<td>Обвързването със задължителен мандат е недействително.</td>
</tr>
<tr>
<td>1851</td>
<td>2139067</td>
<td>16</td>
<td>Народните представители действат въз основа на Конституцията и законите в съответствие със своята съвест и убеждения.</td>
</tr>
<tr>
<td>1852</td>
<td>2139068</td>
<td>23</td>
<td>Народните представители не могат да изпълняват друга държавна служба или да извършват дейност, която според закона е несъвместима с положението на народен представител.</td>
</tr>
</tbody>
</table>
Заседанията на събрание на Народното Събрание открити открия (се).
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

<table>
<thead>
<tr>
<th></th>
<th>tok</th>
<th>lemma</th>
<th>def</th>
<th>Tok/Lemma</th>
<th>Def/Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj</td>
<td>17741</td>
<td>2077</td>
<td>4344</td>
<td>8.541646606</td>
<td>2.091478093</td>
</tr>
<tr>
<td>Adv</td>
<td>7571</td>
<td>533</td>
<td>726</td>
<td>14.20450281</td>
<td>1.362101313</td>
</tr>
<tr>
<td>Noun</td>
<td>49658</td>
<td>3477</td>
<td>8046</td>
<td>14.28185217</td>
<td>2.314063848</td>
</tr>
<tr>
<td>Verb</td>
<td>17977</td>
<td>2058</td>
<td>5163</td>
<td>8.735179786</td>
<td>2.508746356</td>
</tr>
<tr>
<td>Total</td>
<td>92947</td>
<td>8145</td>
<td>18279</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tok/Lemma</td>
<td>11.41154082</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Def/Lemma</td>
<td>2.244198895</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
% selected: висок
% selected: глас.
% def: възвишен
% selected: За звук, глас - който се произвежда от
% selected: 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: отдalenen 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

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPA</td>
<td>head (verb)-adjunct</td>
</tr>
<tr>
<td>VPC</td>
<td>head(verb)-complement</td>
</tr>
<tr>
<td>VPS</td>
<td>head(verb)-subject</td>
</tr>
<tr>
<td>NPA</td>
<td>head(noun)-adjunct</td>
</tr>
<tr>
<td>NPC</td>
<td>head(noun)-complement</td>
</tr>
<tr>
<td>PP</td>
<td>head(preposition)-complement</td>
</tr>
<tr>
<td>PPA</td>
<td>head(preposition)-adjunct</td>
</tr>
<tr>
<td>APC</td>
<td>head(adjective)-complement</td>
</tr>
<tr>
<td>APA</td>
<td>head(adjective)-adjunct</td>
</tr>
<tr>
<td>AdvC</td>
<td>head(adverb)-complement</td>
</tr>
<tr>
<td>AdvA</td>
<td>head(adverb)-adjunct</td>
</tr>
</tbody>
</table>

*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

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

Table 2: Frequency of syntactic Frames
## Ontological Types:

EVENT > PERSON > OBJECT > ARTEFACT > COGNITIVE FACT

<table>
<thead>
<tr>
<th>N</th>
<th>Syntactic Frame</th>
<th>Ontological Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Predicate</td>
<td>No Ontological Restrictions</td>
</tr>
<tr>
<td>2</td>
<td>Predicate – Complement (CLDA)</td>
<td>EVENT</td>
</tr>
<tr>
<td>3</td>
<td>Subject (NP) – Predicate</td>
<td>PERSON</td>
</tr>
<tr>
<td>4</td>
<td>Predicate – Direct Object (NP)</td>
<td>PERSON</td>
</tr>
<tr>
<td>5</td>
<td>Subject (NP) – Predicate – Complement (CLDA)</td>
<td>PERSON - EVENT</td>
</tr>
<tr>
<td>6</td>
<td>Predicate – Direct Object (NP)</td>
<td>OBJECT</td>
</tr>
<tr>
<td>7</td>
<td>Subject (NP) – Predicate – Direct Object (NP) – Indirect Object (PP)</td>
<td>PERSON – ARTEFACT – (for) OBJECT</td>
</tr>
<tr>
<td>8</td>
<td>Subject (NP) – Predicate – Direct Object (NP)</td>
<td>PERSON - PERSON</td>
</tr>
<tr>
<td>9</td>
<td>Predicate – Direct Object (NP)</td>
<td>COGNITIVE FACT</td>
</tr>
<tr>
<td>10</td>
<td>Subject (NP) – Predicate – Direct Object (NP)</td>
<td>PERSON - OBJECT</td>
</tr>
</tbody>
</table>
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

Other publications (2)

Ontology-to-Text Relation Approach


Work done in projects: AsIsKnown, LT4eL, LTfLL
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 multilingualistic 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

Lexicalized Items

Free Phrases
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>
   </termg>
   <def> . . . </def>
</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
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 metonymy 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


Bulgarian Language Technology

Based on
Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. LREC 2004

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)
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

<table>
<thead>
<tr>
<th>Word</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Той</td>
<td><strong>Ppe-os3m</strong></td>
</tr>
<tr>
<td>обаче</td>
<td><em>Cc; Dd</em></td>
</tr>
<tr>
<td>няма</td>
<td><em>Afsi; Vnif-o3s; Vnif-r3s; Vpif-o2s; Vpif-o3s; Vpif-r3s</em></td>
</tr>
<tr>
<td>възможност</td>
<td><em>Ncfsi</em></td>
</tr>
<tr>
<td>да</td>
<td><em>Ta; Tx</em></td>
</tr>
<tr>
<td>следи</td>
<td><em>Ncfpi; Vpif-o2s; Vpif-o3s; Vpif-r3s; Vpif-r3s; Vpitz-2s</em></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>#</th>
<th>Lexicon (source of)</th>
<th>Linguistic Rules (applied to filter):</th>
<th>Beam size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(a) the lexicon features</td>
<td></td>
<td>Sentence-level</td>
</tr>
<tr>
<td>1</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>52.95</td>
</tr>
<tr>
<td>2</td>
<td>–</td>
<td>yes</td>
<td>1</td>
<td>64.50</td>
</tr>
<tr>
<td>3</td>
<td>features</td>
<td>–</td>
<td>1</td>
<td>70.40</td>
</tr>
<tr>
<td>4</td>
<td>features</td>
<td>yes</td>
<td>1</td>
<td>70.30</td>
</tr>
<tr>
<td>5</td>
<td>features</td>
<td>yes, for test only</td>
<td>1</td>
<td>70.40</td>
</tr>
<tr>
<td>6</td>
<td>features</td>
<td>–</td>
<td>1</td>
<td>71.34</td>
</tr>
<tr>
<td>7</td>
<td>features</td>
<td>yes</td>
<td>1</td>
<td>71.69</td>
</tr>
<tr>
<td>8</td>
<td>features</td>
<td>yes</td>
<td>3</td>
<td>71.94</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Gold Tag</th>
<th>Proposed Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Ansi</td>
<td>Dm</td>
</tr>
<tr>
<td>23</td>
<td>Vpitf-r3s</td>
<td>Vnitr-f3r</td>
</tr>
<tr>
<td>16</td>
<td>Npmsi</td>
<td>Npmsi</td>
</tr>
<tr>
<td>14</td>
<td>Vpiif-r3s</td>
<td>Vniiif-r3</td>
</tr>
<tr>
<td>13</td>
<td>Npf3d</td>
<td>Npfsi</td>
</tr>
<tr>
<td>12</td>
<td>Dm</td>
<td>Ansi</td>
</tr>
<tr>
<td>12</td>
<td>V pitcam-smi</td>
<td>V pitcao-smi</td>
</tr>
<tr>
<td>12</td>
<td>Vpptf-r3p</td>
<td>V pitf-r3p</td>
</tr>
<tr>
<td>11</td>
<td>Vpptf-r3s</td>
<td>Vpptf-o3s</td>
</tr>
<tr>
<td>10</td>
<td>Mcmsi</td>
<td>Pfe-os-mi</td>
</tr>
<tr>
<td>10</td>
<td>Ppetas3n</td>
<td>Ppetas3m</td>
</tr>
<tr>
<td>10</td>
<td>Ppetds3f</td>
<td>Psot–3–f</td>
</tr>
<tr>
<td>9</td>
<td>Npnsi</td>
<td>Npnsd</td>
</tr>
<tr>
<td>9</td>
<td>Vpptf-o3s</td>
<td>Vpptf-r3s</td>
</tr>
<tr>
<td>8</td>
<td>Dm</td>
<td>A-pi</td>
</tr>
<tr>
<td>8</td>
<td>Ppxts</td>
<td>Ppxtd</td>
</tr>
<tr>
<td>7</td>
<td>Mcfsi</td>
<td>Pfe-os-fi</td>
</tr>
<tr>
<td>7</td>
<td>Npfsi</td>
<td>Npfsd</td>
</tr>
<tr>
<td>7</td>
<td>Ppetas3m</td>
<td>Ppetas3n</td>
</tr>
<tr>
<td>7</td>
<td>Vnitr-r3s</td>
<td>Vpitf-r3s</td>
</tr>
<tr>
<td>7</td>
<td>V pitcam-p-i</td>
<td>V pitcao-p-i</td>
</tr>
</tbody>
</table>

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;

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
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.
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 10,790 sentences and each test set – 1,198
- 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

<table>
<thead>
<tr>
<th></th>
<th>MLT07</th>
<th>MLT09</th>
<th>MATE01</th>
<th>MST05</th>
<th>Turbo02</th>
<th>training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.900</td>
<td>0.908</td>
<td>0.929</td>
<td>0.911</td>
<td>0.927</td>
<td></td>
<td>gold</td>
</tr>
<tr>
<td>0.881</td>
<td>0.890</td>
<td>0.910</td>
<td>0.890</td>
<td>0.911</td>
<td></td>
<td>BLL tagger</td>
</tr>
<tr>
<td>0.881</td>
<td>0.889</td>
<td>0.908</td>
<td>0.890</td>
<td>0.910</td>
<td></td>
<td>Mate tagger</td>
</tr>
<tr>
<td>0.857</td>
<td>0.865</td>
<td>0.883</td>
<td>0.865</td>
<td>0.883</td>
<td></td>
<td>TreeTagger</td>
</tr>
</tbody>
</table>
## Voting Combination

<table>
<thead>
<tr>
<th>Models</th>
<th>Algorithm</th>
<th>Rank01 Number</th>
<th>Rank02 Sum</th>
<th>Rank03 Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LAS</td>
<td>UAS</td>
<td>LAS</td>
</tr>
<tr>
<td>(1) all models</td>
<td>LocTr</td>
<td>88.55</td>
<td>92.05</td>
<td>88.61</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>88.55</td>
<td>91.96</td>
<td>88.65</td>
</tr>
<tr>
<td>(2) all Mate01 and Turbo02 models</td>
<td>LocTr</td>
<td>87.68</td>
<td>91.38</td>
<td>87.80</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>87.58</td>
<td>91.21</td>
<td>87.82</td>
</tr>
<tr>
<td>(3) best combination</td>
<td>LocTr</td>
<td>88.90</td>
<td>92.34</td>
<td>89.05</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>88.94</td>
<td>92.31</td>
<td>89.14</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>the current node</td>
</tr>
<tr>
<td>WordBefore</td>
<td>the word before the current node</td>
</tr>
<tr>
<td>WordAfter</td>
<td>the word after the current node</td>
</tr>
<tr>
<td>ParentWord</td>
<td>the parent word</td>
</tr>
<tr>
<td>PWordBefore</td>
<td>the word before the parent word</td>
</tr>
<tr>
<td>PWordAfter</td>
<td>the word after the parent word</td>
</tr>
<tr>
<td>SelectedArc</td>
<td>one of the arcs suggested by one of the models for the node</td>
</tr>
<tr>
<td>SelectedTag</td>
<td>one of the arcs suggested by one of the models for the node</td>
</tr>
<tr>
<td>CorrectIncorrect</td>
<td>true or false depending on whether the selected pair is the correct one for the node</td>
</tr>
</tbody>
</table>
Classification vs. Regression

- We first tried RandomForest classification – too coarse:

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>65.2%</td>
<td>73.4%</td>
</tr>
<tr>
<td>False</td>
<td>34.8%</td>
<td>26.6%</td>
</tr>
</tbody>
</table>

- Regression uses a continuous interval of values – fine grained ranking
## Results for Parsing

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>LocTr</td>
<td>89.17</td>
<td>92.46</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>89.23</td>
<td>92.27</td>
</tr>
<tr>
<td>all Mate01 and Turbo02 models</td>
<td>LocTr</td>
<td>88.26</td>
<td>91.81</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>88.32</td>
<td>91.87</td>
</tr>
<tr>
<td>best combination</td>
<td>LocTr</td>
<td>89.76</td>
<td>93.18</td>
</tr>
<tr>
<td></td>
<td>GloTr</td>
<td>89.81</td>
<td>93.22</td>
</tr>
</tbody>
</table>
## Results for Tagging

<table>
<thead>
<tr>
<th>Voting</th>
<th>Rank01</th>
<th>Rank02</th>
<th>Rank03</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Sum</td>
<td>Average</td>
</tr>
<tr>
<td>BLL, Mate, TreeTagger</td>
<td>96.24</td>
<td>96.24</td>
<td>95.22</td>
</tr>
<tr>
<td><strong>MLearning</strong></td>
<td>WMax</td>
<td>WMin</td>
<td>WSum</td>
</tr>
<tr>
<td>BLL, Mate, TreeTagger</td>
<td>96.10</td>
<td>96.20</td>
<td>96.25</td>
</tr>
<tr>
<td>BLL, Mate</td>
<td>96.62</td>
<td>96.59</td>
<td>96.63</td>
</tr>
<tr>
<td>BLL, TreeTagger</td>
<td>95.89</td>
<td>96.08</td>
<td>96.09</td>
</tr>
<tr>
<td>Mate, TreeTagger</td>
<td>95.29</td>
<td>95.40</td>
<td>96.25</td>
</tr>
</tbody>
</table>
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 Analysis for MT transfer
  - (R)MRS structure over dependency parses augmented with ontology and instance information from LOD
Thank you for your attention!