Furthering Natural Language Processing in Bulgaria

Svetla Koeva

Institute for Bulgarian, Bulgaria

svetlal@dcl.bas.bg

META-FORUM
Budapest, Hungary, 2011-06-27-28

Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

Co-funded by the ICT PSP Programme of the European Commission through the contract CESAR, grant agreement no.: 271022.

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

Furthering NLP in Bulgaria

General facts

- Republic of Bulgaria
- Area - 110, 993. 6 km2
- Population - 7 351 633
- Bulgarian -
  9 million
  native speakers

General facts

- The official alphabet is Cyrillic.
- Официалната азбука е кирилица.
- Cyrillic became the third official alphabet of the European Union, following the Latin and Greek alphabets.

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

Research

Furthering NLP in Bulgaria

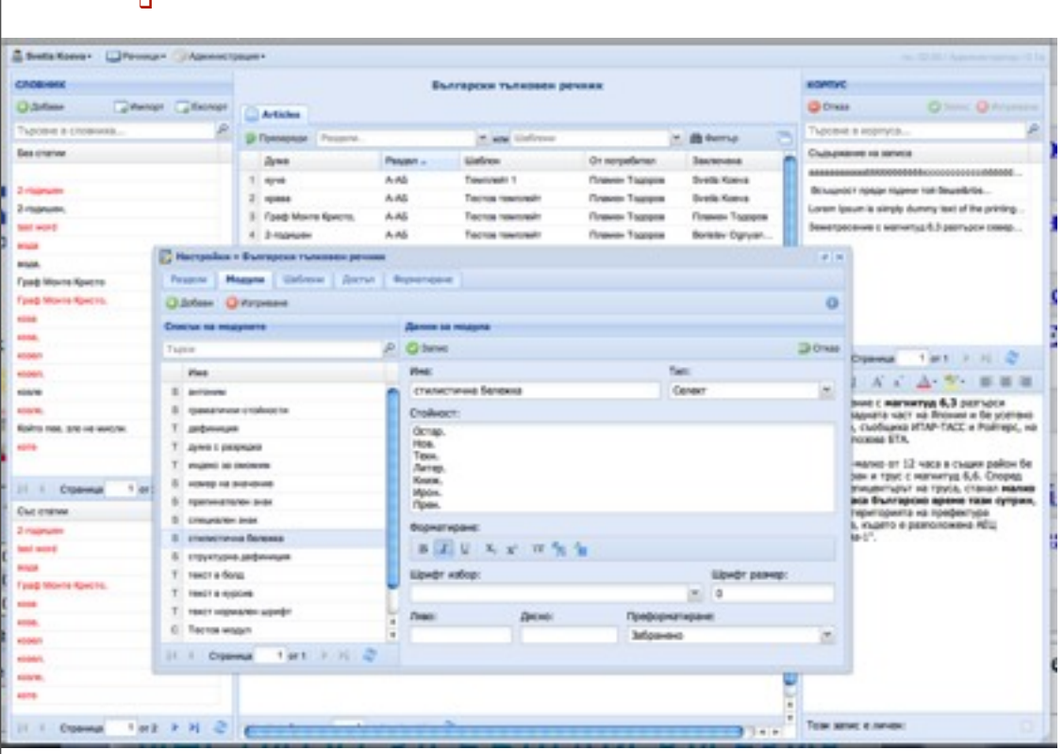
BLARK

- Over the past decade a number of important language resources and tools have been developed.

Furthering NLP in Bulgaria

BLARK

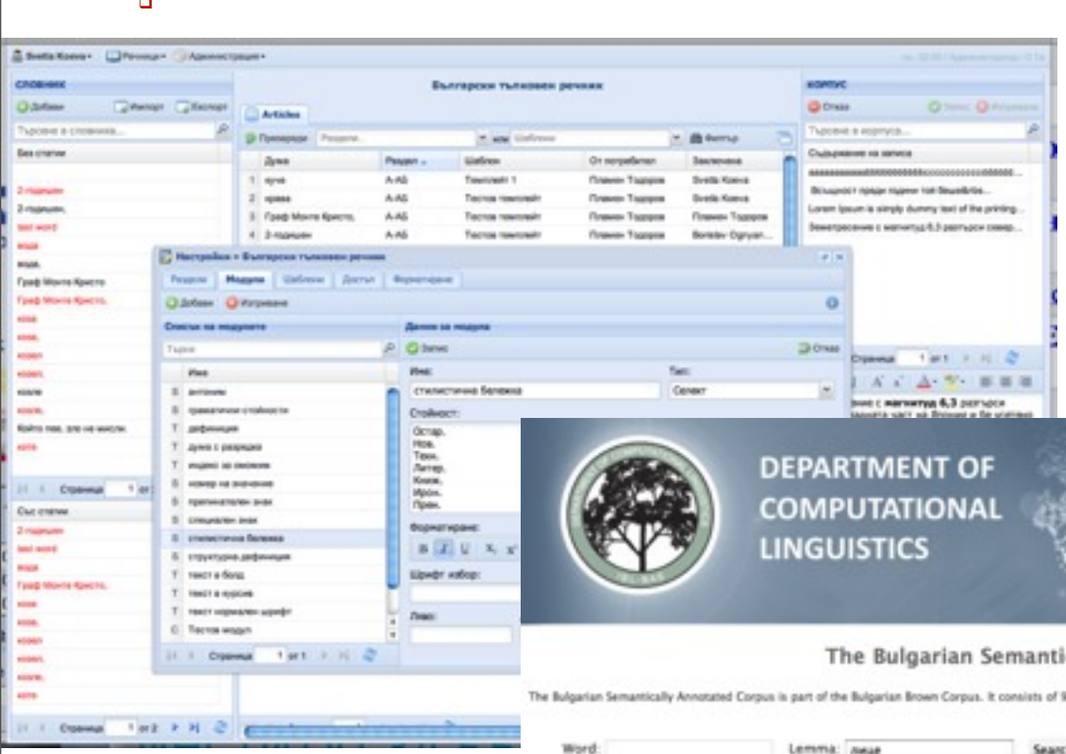
- Over the past decade a number of important language resources and tools have been developed.



Furthering NLP in Bulgaria

BLARK

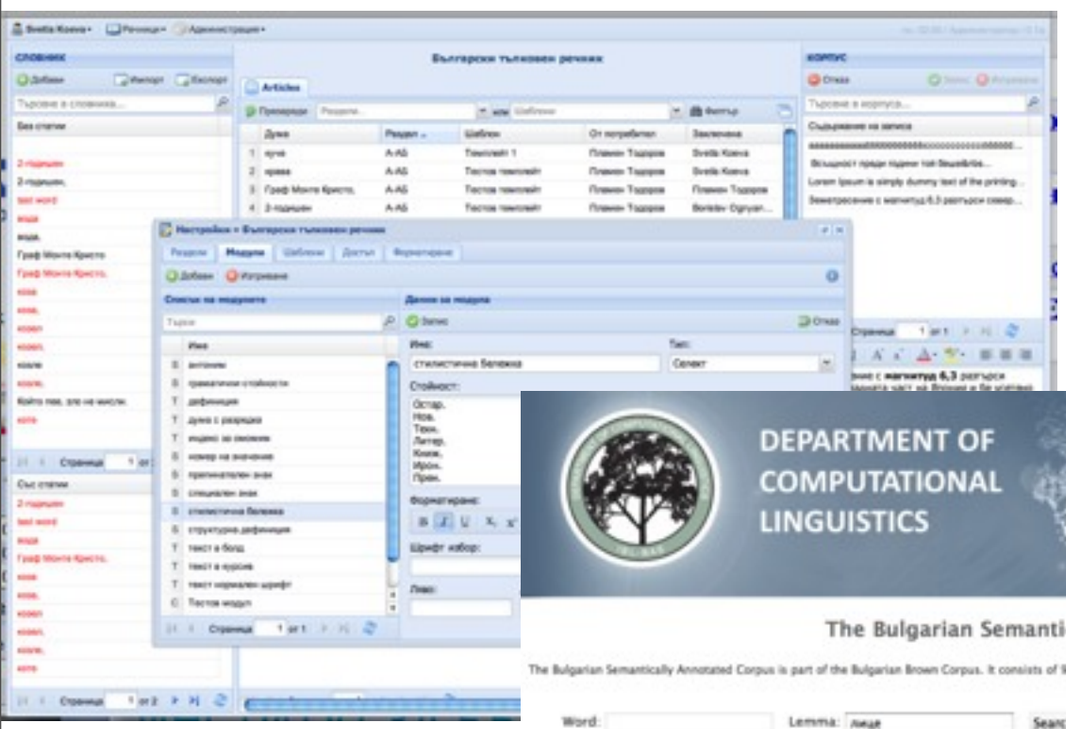
- Over the past decade a number of important language resources and tools have been developed.

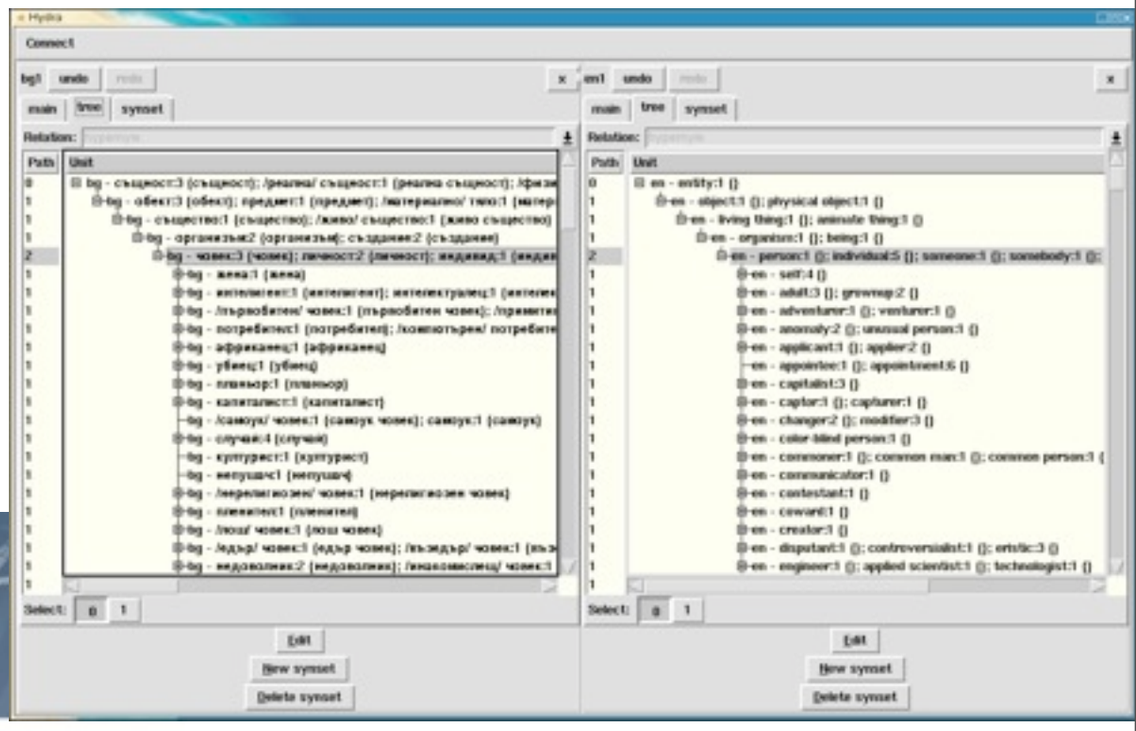


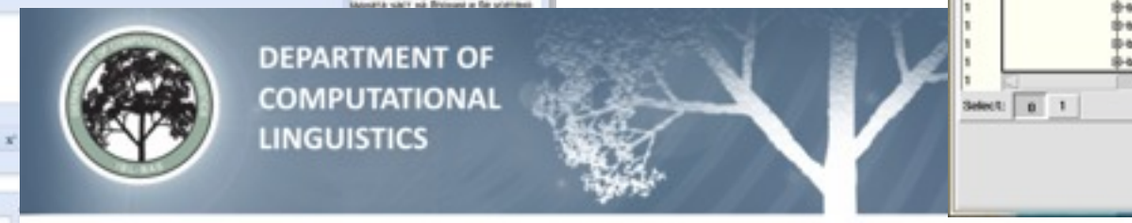
Furthering NLP in Bulgaria

BLARK

- Over the past decade a number of important language resources and tools have been developed.







The Bulgarian Semantically Annotated Corpus

The Bulgarian Semantically Annotated Corpus is part of the Bulgarian Brown Corpus. It consists of 95119 lexical units, annotated with the most appropriate synonymous set from the Bulgarian wordnet.

Word: Lemma: лице Search

Results:

- **Occurences: 43, sense definition: предната по-плоска част на главата на човек, ограничена от челото и челюстите, заедно с разположените върху нея органи - уста, очи, нос**

  Събитият провал изписа широки усмивки на удовлетворение по **лицата** на хората, които през последните две седмици открито се обявяха срещу регистрацията на партията по принцип и срещу начина, по който организаторите подготвят несъстоялото се събитие.

- **Occurences: 16, sense definition: представител на едноименния и единствен вид на същото семейство висши бозайници (Hominidae), различаващ се от останалите животни по силно развития мозък, съзнание, абстрактно мислене, членоразделна реч; движи се с изправено тяло, произвежда оръдия на труда и други артефакти**

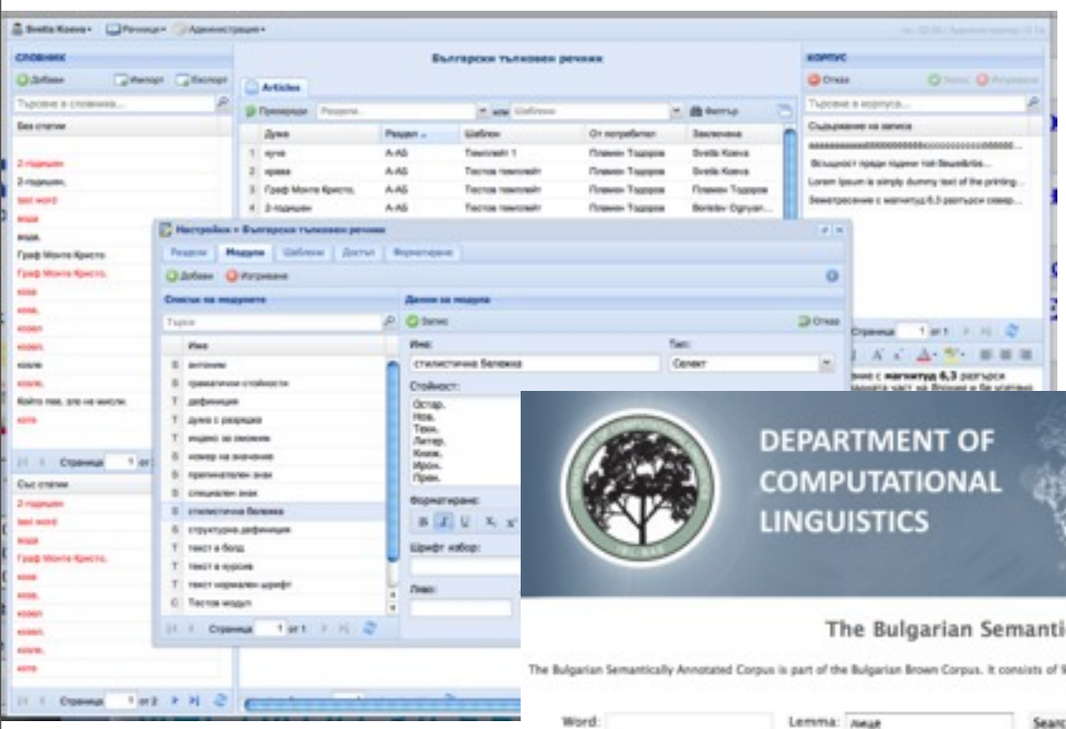
  - Бъдете спокоен, господин полковник, в мое **лице** ще имате добър водач.

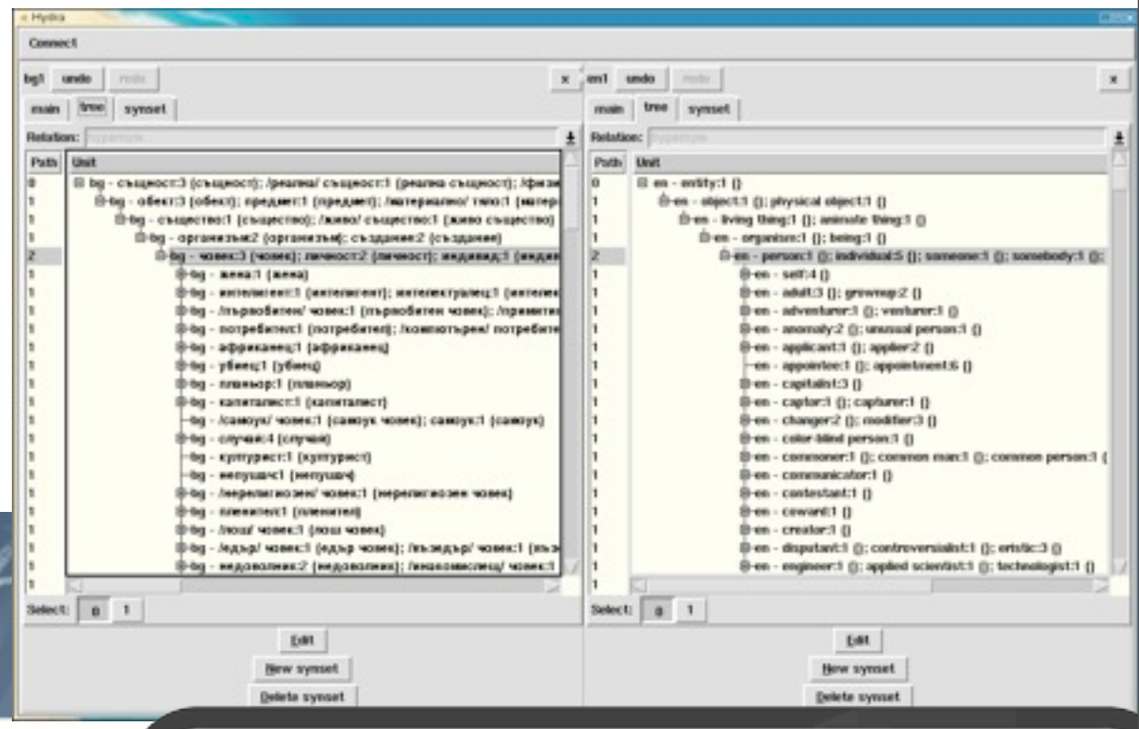
- **Occurences: 4, sense definition: метонимично название за отделен човек**

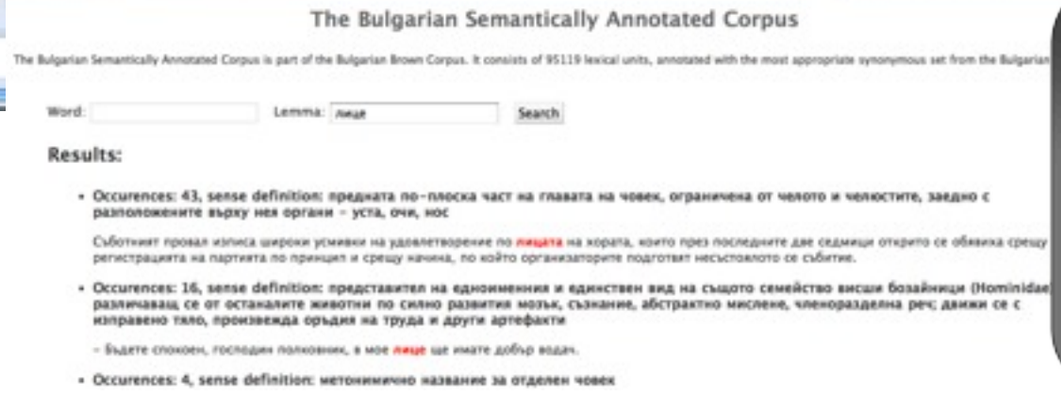
Furthering NLP in Bulgaria

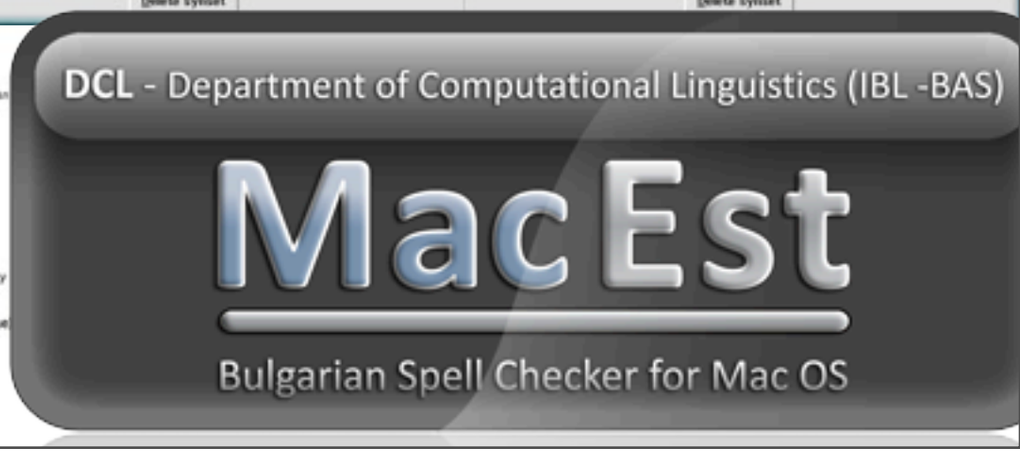
BLARK

- Over the past decade a number of important language resources and tools have been developed.





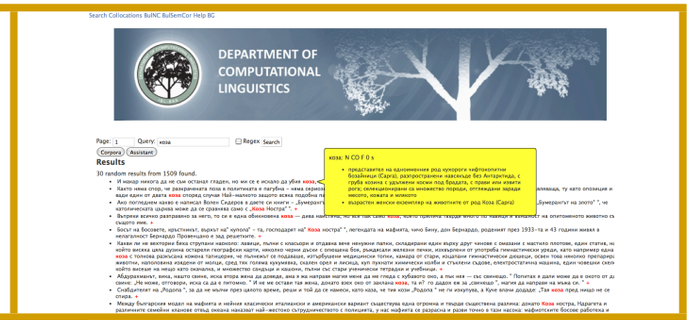




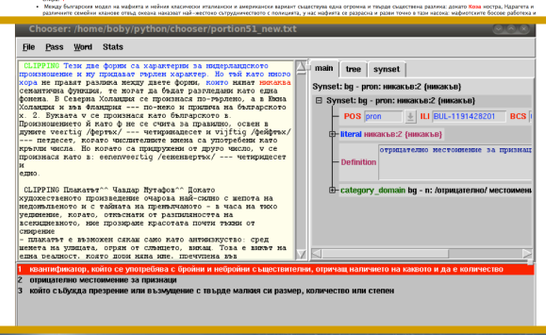
Furthering NLP in Bulgaria

BLARK

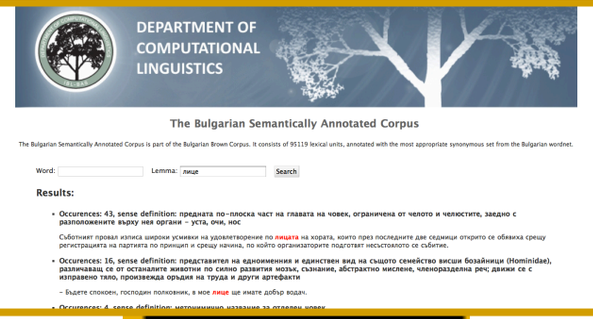
Bulgarian National Corpus - app. 500M words



Bulgarian POS-annotated Corpus



Bulgarian Sense-annotated Corpus



Dependency part of BulTreeBank

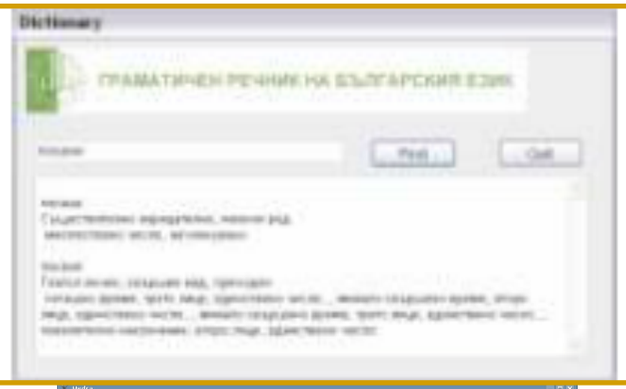
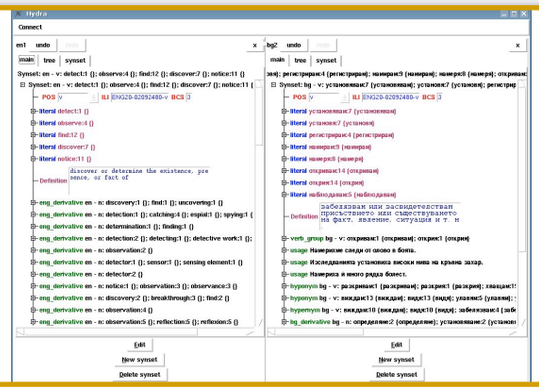
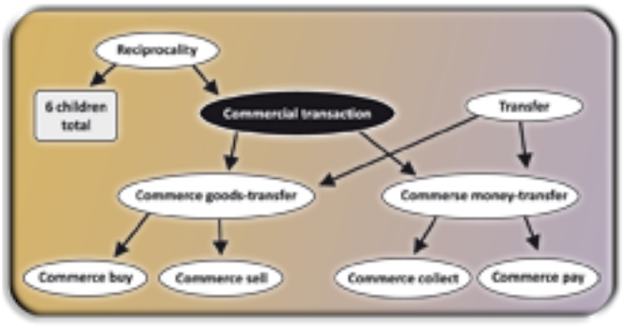
Furthering NLP in Bulgaria

BLARK

- SEE-ERA.net Administrative and Literally Corpus
- Bilingual collection of cultural texts in Greek and Bulgarian
- Bulgarian-Polish-Lithuanian Corpus
- Bulgarian-English-X language parallel corpus - app. 100M words for Bulgarian
- ...

Furthering NLP in Bulgaria

BLARK

Several large inflectional dictionaries	
Bulgarian WordNet	
Bulgarian FrameNet	

Furthering NLP in Bulgaria

Companies

- Development of tools and solutions based on semantic technologies
- Ontology design
- Data integration, management and publishing

Furthering NLP in Bulgaria

Companies

cubist — Your Business Intelligence	LarKC — Large Knowledge Collider	NoTube
insemtives — Incentives for Semantics	MOLTO	RENDER — Reflecting Knowledge Diversity
khresmoi — Medical Information Analysis & Retrieval		SOA All — Semantics, Web 2.0, Web Services, Context

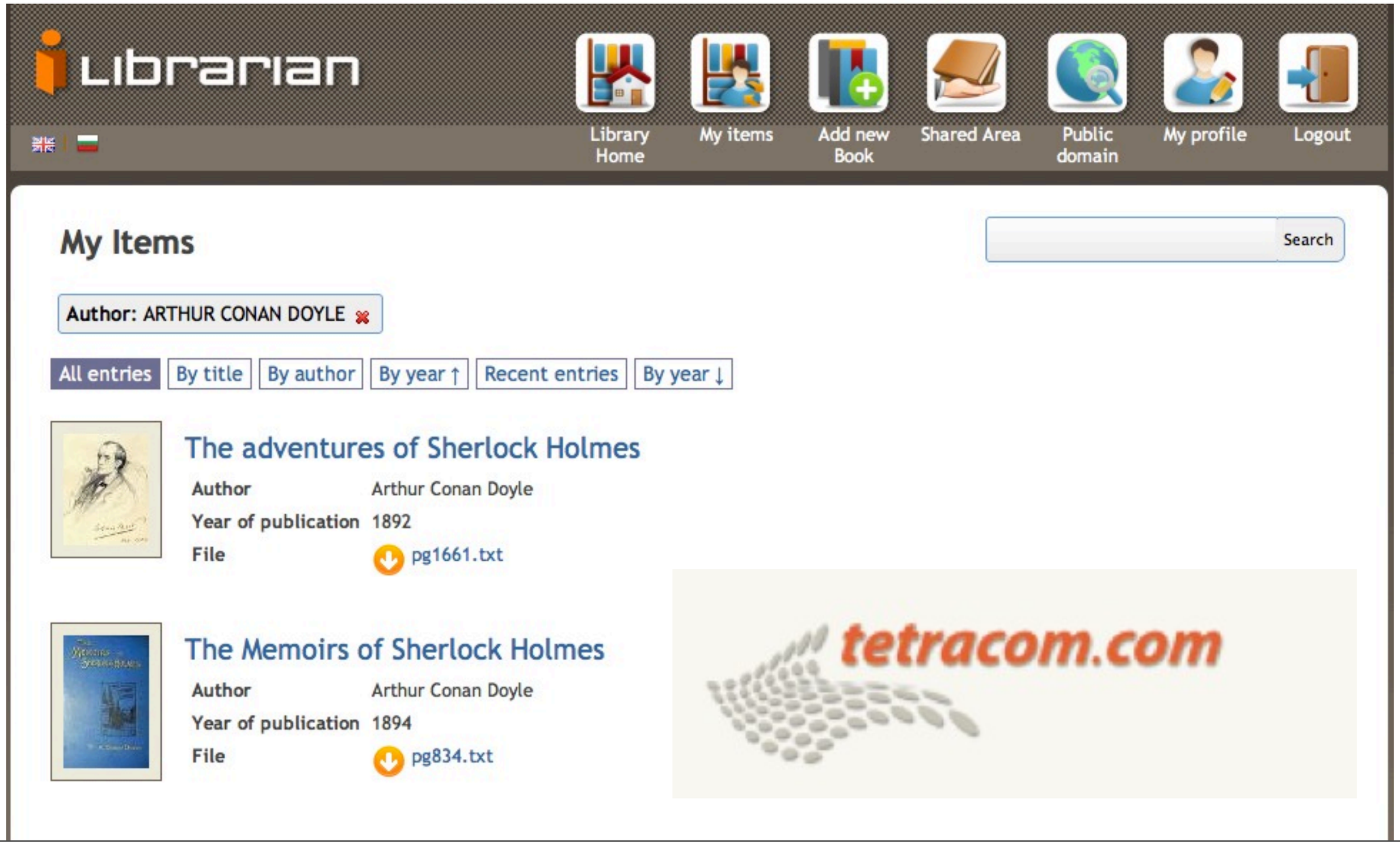
Furthering NLP in Bulgaria

Companies

- Web applications (dynamic web content)
- Content Management Systems (CMS)
- Tools for web site content management
- Multilingual tools and services for natural language processing

Furthering NLP in Bulgaria

Companies



Furthering NLP in Bulgaria

Companies

SkyCode

- WebTrance - a translation software package from English, French, German, Spanish, Italian and Turkish to Bulgarian and vice versa.
- SkyCode is one of the partners of *iTranslate4*.

Furthering NLP in Bulgaria

Approximate status

Technology	Median
Tokenization, Morphology	4
Parsing	3
Information Retrieval	2
Speech Synthesis	2
Text semantics	2
Information extraction	2
Summarization, QA	2
Machine translation	2
Language generation	1

Resources	Median
Reference Corpora	4
Thesauri, WordNets	4
Lexicons, Terminologies	3
Semantic corpora	3
Parallel Corpora, TM	2
Syntax-Corpora	2
Discourse-Corpora	1
Multimedia/multimodal data	1

http://www.meta-net.eu

Furthering NLP in Bulgaria

Approximate status

Technology	Median
Tokenization, Morphology	4
Parsing	3
Information Retrieval	2
Speech Synthesis	2
Text semantics	2
Information extraction	2
Summarization, QA	2
Machine translation	2
Language generation	1

Resources	Median
Reference Corpora	4
Thesauri, WordNets	4
Lexicons, Terminologies	3
Semantic corpora	3
Parallel Corpora, TM	2
Syntax-Corpora	2
Discourse-Corpora	1
Multimedia/multimodal data	1

http://www.meta-net.eu

Furthering NLP in Bulgaria

Approximate status

Technology	Median
Tokenization, Morphology	4
Parsing	3
Information Retrieval	2
Speech Synthesis	2
Text semantics	2
Information extraction	2
Summarization, QA	2
Machine translation	2
Language generation	1

Resources	Median
Reference Corpora	4
Thesauri, WordNets	4
Lexicons, Terminologies	3
Semantic corpora	3
Parallel Corpora, TM	2
Syntax-Corpora	2
Discourse-Corpora	1
Multimedia/multimodal data	1

http://www.meta-net.eu

Furthering NLP in Bulgaria

Approximate status

Technology	Median
Tokenization, Morphology	4
Parsing	3
Information Retrieval	2
Speech Synthesis	2
Text semantics	2
Information extraction	2
Summarization, QA	2
Machine translation	2
Language generation	1

Resources	Median
Reference Corpora	4
Thesauri, WordNets	4
Lexicons, Terminologies	3
Semantic corpora	3
Parallel Corpora, TM	2
Syntax-Corpora	2
Discourse-Corpora	1
Multimedia/multimodal data	1

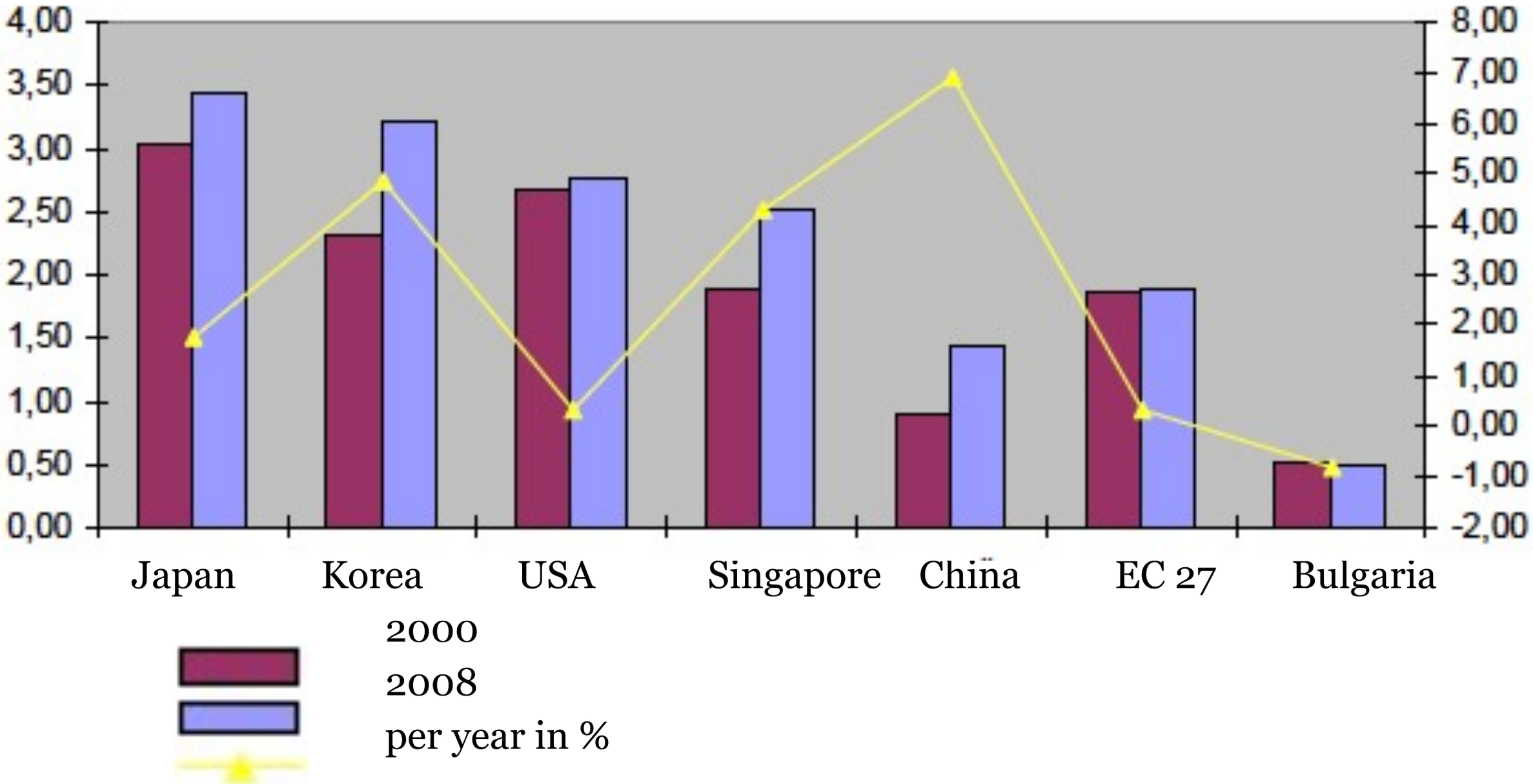
Furthering NLP in Bulgaria

Approximate status

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Technology	2	2	2.5	2.5	2	2	2.5
Resources	2	2.5	3	3.5	2.5	2.5	2.5
Total	2	2	2.5	3	2	2	2.5

Furthering NLP in Bulgaria

State contribution to R&D



Furthering NLP in Bulgaria

State contribution to R&D

- Strategic research agenda:
- Cultural-historical heritage - language being a central part of it
- ICT as a horizontal instrument

Furthering NLP in Bulgaria

European dimensions

- **META, Multilingual Europe Technology Alliance**

- Institute for Bulgarian, BAS, a member of
- Institute for Literature, BAS
- Institute of Information and Communication Technologies, BAS

- Sofia University St. Kliment Ohridski
- University of Plovdiv

- Ontotext, Bulgaria
- Musala Soft, Bulgaria
- Tetracom Interactive Solutions, Bulgaria
- TransGlobe International Ltd., Bulgaria

Conclusions

- Several factors are mutually related for the success:
  - clear formulation of target goals and strategies for their accomplishment
  - stable financing
  - effective management of the resources
  - beneficial relations between education - research - business - end users
  - networking
- META-NET as a concerted, substantial, continent-wide effort in language technology research and engineering is relevant for all of these factors.

- Thank you very much for your attention.