ONLINE SERVICES

Bulgarian National Corpus
http://search.dcl.bas.bg

Bulgarian WordNet
http://dcl.bas.bg/bulnet/

Bulgarian Proofing Tools
http://dcl.bas.bg/est/

MediaTalk
http://www.mtalk.eu

CONTACT DETAILS

Department of Computational Linguistics
Institute for Bulgarian Language Prof. Lyubomir Andreychin
Bulgarian Academy of Sciences

52 Shipchenski prohod blvd., bldg. 17
Sofia, 1113
Bulgaria

tel.: +359 2 / 979 2969, +359 2 / 979 2971
tel./fax: +359 2 / 872 23 02
e-mail: dcl@dcl.bas.bg

DEPARTMENT OF COMPUTATIONAL LINGUISTICS

- Institute for Bulgarian Language
- Bulgarian Academy of Sciences

DEPARTMENT OF
COMPUTATIONAL
LINGUISTICS

Bulgarian Academy of Sciences

http://dcl.bas.bg

Learn from yesterday, live for today, hope for tomorrow. The important thing is to not stop questioning.

Albert Einstein, Relativity: The Special and the General Theory

Science is not only compatible with spirituality; it is a profound source of spirituality.

Carl Sagan, The Demon-Haunted World: Science as a Candle in the Dark

The opposite of a correct statement is a false statement. But the opposite of a profound truth may well be another profound truth.

Niels Bohr

European society is multilingual: the diversity of its cultural heritage is an asset and an opportunity.

Language barriers must be overcome: language technology is a key enabler which will help to solve this problem.

META-NET
Strategic Research Agenda for
Multilingual Europe 2020

DEPARTMENT OF COMPUTATIONAL LINGUISTICS

GENERAL

The Department of Computational Linguistics (DCL) is a leading centre for theoretical and applied research in computational linguistics and language technologies in Bulgaria and the region.
The Department of Computational Linguistics is:

- founded in 1994 as Department of Computer Modelling of Bulgarian with the Institute for Bulgarian Language at the Bulgarian Academy of Sciences;
- highly interdisciplinary, with close collaborations not only within but outside the country;
- a partner in many international projects under the EU framework and other scientific programmes;
- a member of research infrastructures for language resources, tools and technologies;
- an active knowledge dissemination centre.

TEAM

Highly-qualified researchers - computational linguists, software engineers, mathematicians and logicians. The team includes full-time researchers, part-time associates and Ph. D. students.

PRIORITIES

- Creating mono- and multilingual language resources and tools based on the state-of-the-art in the field
- Bringing innovations to linguistic research and language technologies
- Building research infrastructures and networks for strategic scientific cooperation
- Promoting the achievements of Computational linguistics and the societal benefits of language technologies to the public

RESEARCH

The major fields of research carried out at the Department of Computational Linguistics include:

- theoretical issues in the formal description of language;
- morphological, syntactic and semantic analysis;
- morphological, syntactic, and semantic disambiguation;
- electronic dictionaries and lexical-semantic networks;
- ontologies and semantic relations;
- document categorisation and text clustering;
- document summarisation;
- information extraction and information retrieval;
- machine translation.

EDUCATION

The Department of Computational Linguistics offers high quality education in theoretical and computational linguistics within:

- Ph.D. programme in Computational linguistics;
- student apprenticeships and traineeships;
- postdoctoral fellowships.

Members of the DCL team are among the founders and lecturers of the Master's programme in Computational linguistics at Sofia University.

PARTNERS

The Department of Computational Linguistics has established a long-standing cooperation with leading research and academic institutions in Bulgaria: Sofia University, Plovdiv University, the Institute of Mathematics and Informatics, the Institute of Information and Communication Technologies, Tetracom LTD, SkyCode, and many others.

The Department of Computational Linguistics has actively collaborated with research centres from Croatia, the Czech Republic, France, Germany, Greece, Hungary, Norway, Poland, Romania, Serbia, Slovakia, Turkey, the UK, and other countries under large European projects, such as BalkaNet, CESAR, and ATLAS, as well as within the Network of Excellence META-NET, and other initiatives.

CONFERENCES

The Department of Computational Linguistics has organised several international conferences in computational linguistics and formal description of natural languages, and has hosted numerous lectures and seminars by renowned scientists.

51st ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS
The 51st Annual Meeting of The Association for Computational Linguistics
ACL 2013
ACL 2013

The DCL team was the local organiser of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013).

MEMBERSHIP

ACL - Association for Computational Linguistics

BACL – Bulgarian Association for Computational Linguistics

META - Multilingual Europe Technology Alliance

FoLLI - Association for Logic, Language and Information

The Institute for Bulgarian Language is a member of META-NET, a Network of Excellence consisting of 60 research centres from 34 countries. META-NET is dedicated to building the technological foundations of a multilingual European information society.

PROJECTS

CESAR: CEntral and South-East EuropeAn Resources (2011-2013)

The CESAR PSP ICT EU project, in close harmony with META-NET, enhances, upgrades, standardises, cross-links and makes available a comprehensive set of **language resources and tools** for European languages: language data, such as written and spoken corpora; language-related data, including and/or associated to other media and modalities; language processing and annotation tools and technologies; services through the use of language processing tools and technologies; service workflows by combining and orchestrating interoperable services.

ATLAS: Applied Technology for Language-Aided CMS (2010 - 2013)

The ATLAS PSP ICT EU project facilitates the multilingual web content development and management, in particular the authoring, versioning and maintenance of multilingual web sites. ATLAS combines a language processing framework with a content management component (i-Publisher) used for creating, running and managing dynamic content-driven websites. The language processing framework provides automatic annotation of important words - noun phrases and named entities, automatic categorisation of documents, summary generation, and machine translation of a summary of documents for six languages.

BulNC: Bulgarian National Corpus (2009-2016)

The project extends the Bulgarian National Corpus with new corpus samples, both monolingual and parallel, and new levels of annotation and metadata description. The approach integrates the best practices of corpus compilation with the potential of the latest technologies allowing effective mining and collection of documents published on the internet, automatic metadata description and linguistic annotation of large amounts of data, automatic alignment of parallel texts in different languages.

Language Resources and Processing Tools (2011-2016)

The project aims at expanding and validating the Bulgarian WordNet, and further developing the Bulgarian FrameNet. The project also involves implementation of tools for creation, editing and visualisation of the Bulgarian WordNet and the Bulgarian FrameNet consistent with recent theories. The integrated suite of natural language processing tools for Bulgarian, including tokenisation, part-of-speech tagging, named entity recognition, is being further expanded with tools for sense disambiguation, parsing, and coreference resolution.

Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics (2012-2014)

The project aims at applying and upgrading the knowledge and skills acquired by postgraduate and postdoctoral students and young researchers in formal training programmes toward the advancement of their research and teaching competences and the promotion of their professional careers in the field of Computational linguistics.

PARSEME
PARSing and Multiword Expressions (2013-2016)

ENeL
European Network of e-Lexicography (2013-2016)

RESOURCES

DICTIONARIES

Different types of dictionaries developed in DCL (dictionaries of general and specialised lexis, grammar dictionaries of simple and compound words, dictionaries of proper names, abbreviations, frequency dictionaries, etc.) are widely used in various research and applied tasks - spelling and grammar checking of simple and compound words, tagging and lemmatisation, phrase and named entity extraction, etc.

WORDNET

The Bulgarian WordNet (BulNet) is a large lexical-semantic network of Bulgarian. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (more than 53,000 synsets), each expressing a distinct concept. In addition, the Bulgarian WordNet contains pronouns, prepositions, conjunctions, particles and interjections. Synsets are interlinked by means of conceptual-semantic and lexical relations. Each Bulgarian synset is linked to its counterpart in the Princeton WordNet 3.0. The structure and organisation of information in WordNet makes it a useful resourse for computational linguistics and natural language processing.

CORPORA

The **Bulgarian PoS-annotated Corpus** contains 174,697 lexical units (words and multiword expressions), lemmatised and manually assigned part-of-speech tags and unambiguous grammatical information.
The **Bulgarian Sense-Annotated Corpus** contains over 99,480 lexical units (words and multiword expressions), manually assigned a unique sense which best corresponds to the meaning in context.
The **Bulgarian-English Sentence- and Clause-Aligned Corpus** consists of 176,397 tokens for Bulgarian and 190,468 for English. The parallel clauses in the English and the Bulgarian texts are manually aligned within the corresponding aligned sentences resulting in 24,654 parallel clause beads.

BULGARIAN NATIONAL CORPUS (BULNC)

Content: reflects the state of the Bulgarian language (mainly in its written form) from the middle of the XX c. (1945) until the present and contains multilingual/parallel content in Bulgarian and other languages in recent years (after 2000).

Type: written and spoken language; multilingual; general, with a number of specialised subcorpora; supplied with metadata description and multi-level linguistic annotation.

Language(s): Bulgarian and 47 other languages (in the Bulgarian-X-Language corpora).

Size: more than 240,000 text samples distributed in 9 general text categories, 130 domains and 100 genres. Overall size: 1.2 billion words for Bulgarian, and 5.4 billion words total for all languages.

Annotation: monolingual - tokenisation, sentence splitting, POS tagging, lemmatisation, WordNet sense annotation; multilingual - sentence and clause alignment for parts of the corpus; detailed metadata description.

Access:

- free online access through the web search system of the Bulgarian National Corpus;
- free download of copyright-free subcorpora;
- free download of a collection of frequency dictionaries extracted from the BulNC.

ЕЗИКОВЕДСКИ ИЗСЛЕДВАНИЯ
В ЧЕСТ НА 75-ГОДИШНИНАТА ОТ
РОЖДЕНИЕТО НА СТ.Н.С. I СТ. Д.Ф.Н.
ЙОРДАН ПЕНЧЕВ ПЕНЧЕВ
ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК
ГАРСКИЯТ СЕМАНТИЧНО
ТИРАН КОРПУС
Verba magistri
АРГУМЕНТНА
СТРУКТУРА
ПРОБЛЕМИ НА ПРОСТОТО
И СЛОЖНОТО ИЗРЕЧЕНИЕ
Formaliser les langues avec l'ordinateur
De INTEX à NooJ
ИНСТИТУТ ЗА БЪЛГАРСКИ ЕЗИК
БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
Българският
ФРЕЙМНЕТ
2010
София 2010
Нови изследвания по генеративен синтаксис на българския език
КРАТКА ПРАКТИЧЕСКА
ГРАМАТИКА
НА БЪЛГАРСКИЯ ЕЗИК

• Focused Solutions •

BASIC NLP PROGRAMS

Web-Based Infrastructure for Data Processing of Bulgarian:

- Sentence segmentation;
- Tokenisation;
- POS tagging and grammatical annotation;
- Lemmatisation;
- NP and NE extraction.

NLP IN THE LINGUISTIC RESEARCH AND EDUCATION

Chooser - an OS-independent multi-functional system for linguistic annotation adaptable to different linguistic levels and different annotation schemata.

Hydra - an OS-independent system designed for wordnet development, validation and exploration. The program enables users to browse and edit any number of monolingual wordnets simultaneously.

Corpora Search Engine - an online query engine of collocations and concordances with advanced queries and regular expressions designed to support monolingual and parallel corpora in a uniform way.

LexIt - an online system for developing, editing and viewing various types of dictionaries.

NLP IN THE MODERN SOCIETY

Proofing tools

The Est family includes applications for detecting errors in Bulgarian texts and generating the most appropriate replacement suggestions (spelling and grammar checking) as well as for improving the quality of working with Bulgarian texts (looking up words in a thesaurus and multilingual dictionaries).

WebEst - Online services for spelling and grammar checking.

WInEst - Bulgarian Spelling Checker for Windows.

MacEst - Bulgarian Spelling Checker for Mac OS X.

MediaTalk is a powerful tool for data analyses, which supersedes the increasingly obsolete search engines. It performs linguistic analysis on customer data pool to identify key phrases representing users' interests and searches the media flow to find similarity on context rather than in words. Further linguistic analyses discover relations between people, places, topics, in order to facilitate decision making, planning and management.

META SHARE

META-SHARE is a multi-layer infrastructure that makes available quality documented language resources and related metadata over the network.

The Department of Computational Linguistics maintains one of the META-SHARE network nodes. **http://metashare.ibl.bas.bg**