# Bulgarian National Corpus

**Svetla Koeva, Ivelina Stoyanova**

Institute for Bulgarian Language, Bulgarian Academy of Sciences | {svetla,iva}@dcl.bas.bg

12th Workshop on the Challenges in the Management of Large Corpora (CMLC-12)

Language Resources and Evaluation Conference ◆ Palma de Mallorca, Spain ◆ 11-16 May 2026

## Motivation and objectives

Since its establishment in 2009, the main features of BulNC include:

- **Diversity of data** in terms of registers, domains, time periods, authors, **etc.**
- **Extensive metadata description.**
- **Linguistic integrity.**

Key areas of progress in recent years include:

- Compilation and use of large volumes of multilingual and, to some extent, multimodal data.
- Moving beyond simple corpus search and analysis to customised linguistic analysis, such as defining words, tracking usage, detecting semantic shifts, and extracting usage examples.
- Serving as clean data for LLM pre-training and fine-tuning.

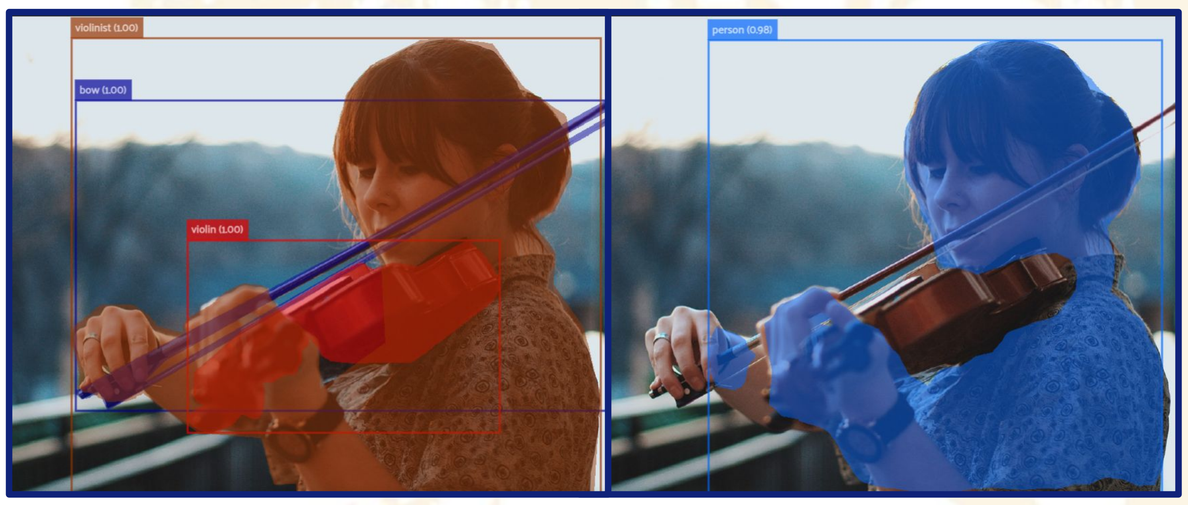
These efforts have led to the development of **IfGPT Dataset** – a large BulNC-based dataset with a special focus on:

- efficient management of large text data,
- various levels of annotation (including parallel aligned corpora),
- multimodal corpora suitable for a wide range of NLP and AI applications.

## Multimodal data

Multilingual Image Corpus (MIC21):

- **Images**. 21K images, copyright-free, drawn from four thematic domains (Sport, Transport, Art, and Security) across 130 subdomains. Images are supplied with metadata.
- **Annotations**. A total of 203K annotated objects, first automatically generated polygon annotations using Detectron2 model which are then manually corrected.
- Object classes are organised in an **Ontology of Visual Objects**, with some classes and relations inherited from WordNet and additional ones added where WordNet lacks coverage.
  The ontology supports extracting object relationships, building datasets at varying levels of granularity, etc.
- **Multilingual support**. Object labels are linked to synonyms, definitions, and usage examples in **25 languages**, drawn from the Extended Open Multilingual Wordnet, BabelNet, and machine translation.




## IfGPT Dataset

| Source | # texts | # tokens | Licenses |
|---|---|---|---|
| MARCELL | 25K | 45M | PD |
| CURLICAT | 113K | 35M | CC |
| BulNC Admin | 17K | 79M | PD |
| BulNC Wikipedia | 89K | 41M | CC/GNU |
| BulNC Subtitles | 146K | 27M | OPUS |
| BG News | 2,116K | 601M | various |
| EN News | 5,961K | 3,324M | various |
| BG Internet | 66K | 289K | various |
| EN Internet | 45K | 8,144M | various |
| News up to 1990 | 5,544K | 271M | various |
| Periodicals up to 1990 | 25K | 30M | various |
| New periodicals | 4,119K | 4,378M | various |
| Books | 22K | 630M | various |

## IfGPT dataset processing pipeline

A pipeline for the integration of resources and the preparation of data tailored for language technologies and LLMs:

- **File handling module** — manages files in plain text, JSONL, and CSV formats using the adopted metadata schema.
- **Dataset quality maintenance module** — covers string manipulation, data cleaning, and error handling, supporting deduplication, PII identification and labelling, and bias detection.
- **Metadata extraction module** — obtains metadata from document sources and content, and provides appropriate metadata descriptions.
- **Annotation module** — introduces linguistic annotation in CoNLL-U Plus format.
- **Dataset construction module** — creates subdatasets for specific purposes based on extensive metadata.
- **Search module** — provides an online interface for browsing and selecting metadata values; outputs either a newly constructed subdataset or download links for relevant parts.

## Access

- Web search interface to the Bulgarian National Corpus for extracting examples: https://search.dcl.bas.bg/. It supports complex linguistic queries involving different levels of annotation.
- IfGPT Dataset search interface: https://ifgpt.dcl.bas.bg/ifgpt-dataset/. It allows to browse and filter the large collection of clean, deduplicated Bulgarian text documents by several criteria – type of licence, domain, time period, keywords.
- Multilingual Image Corpus (MIC21): https://dcl.bas.bg/MIC-21/.

The work is part of the project *Infrastructure for Fine-tuning Pre-trained Large Language Models*, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.
https://ifgpt.dcl.bas.bg/

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan
of the Republic of Bulgaria