IfGPT: A Dataset in Bulgarian for Large Language Models

Svetla Koeva, Ivelina Stoyanova, Jordan Kralev
Institute for Bulgarian Language, Bulgarian Academy of Sciences
svetla@dcl.bas.bg | iva@dcl.bas.bg | jkralev@dcl.bas.bg

Advancing NLP for Low-Resource Languages (LowResNLP 2025) @ RANLP 2025

Motivation and objective

The main issues related to low-resource languages such as Bulgarian:

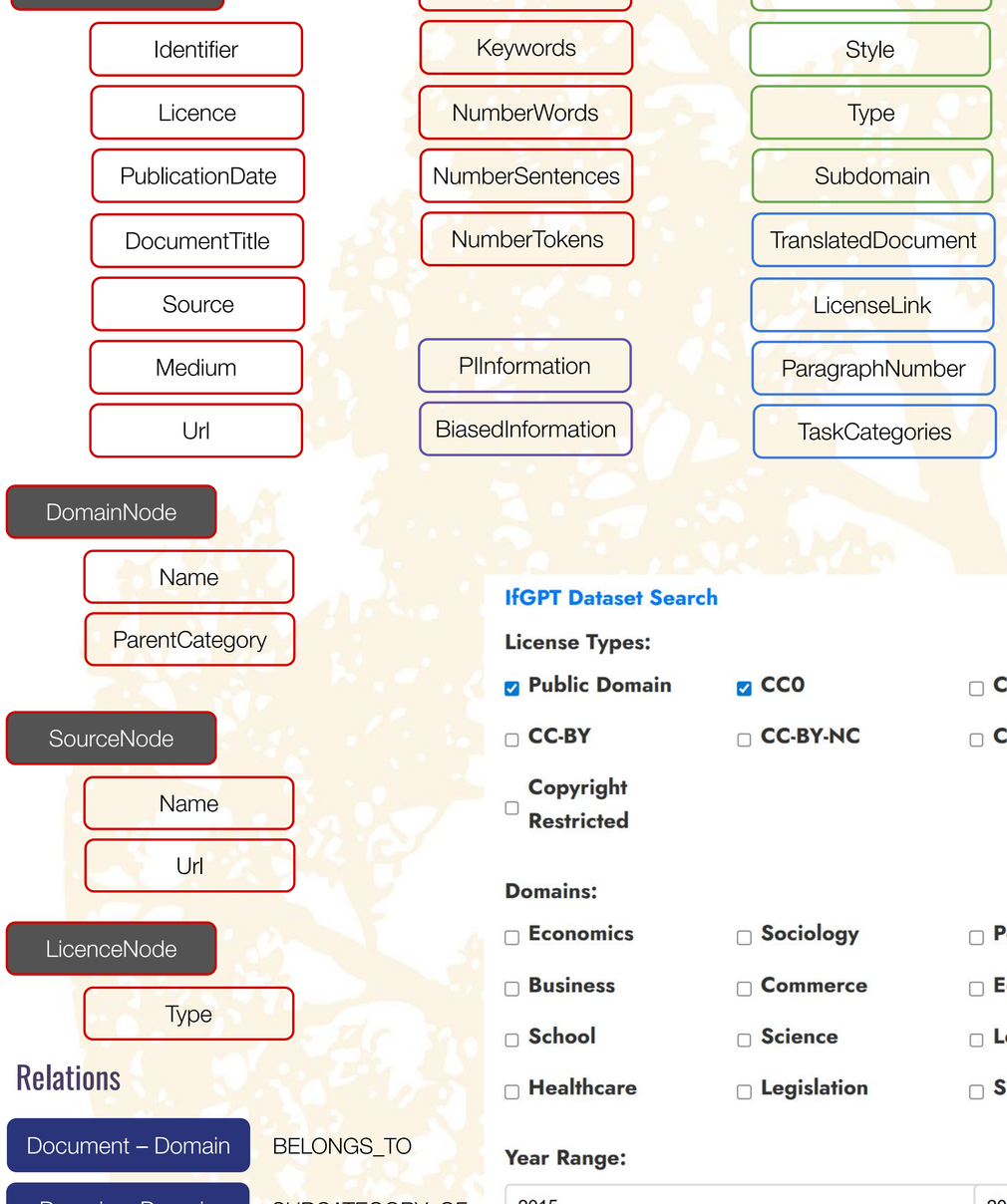
- **Data scarcity** – few sources for compiling large datasets for pre-training and fine-tuning LLMs.
- **Copyright restrictions** – limited datasets with no copyright issues, or legislation is unclear.
- **Quality of the data** – freely accessible data is often noisy and inhomogeneous.

The **main objective** is to compile a **large dataset for Bulgarian**:

★combining existing corpora with newly compiled datasets,
★ensuring the texts are clean, deduplicated and of high-quality,
★supplied with extensive metadata.

IfGPT: Metadata and dataset management

★ Metadata are stored in a **Neo4J graph database**.
★ **Document nodes** are described with a number of properties such as Source, Author, Licence, Domain, etc.



Sources of Bulgarian data for IfGPT

- **Existing datasets:**
  - Bulgarian National Corpus (420 mln. tokens); each text labelled with metadata.
  - General News in Bulgarian (600 mln. tokens); news from different thematic domains.
  - Bulgarian CURLICAT (35 mln. tokens); texts from various sources and domains: Culture, Education, EU, etc.
  - Bulgarian MARCELL (45 mln. tokens); legal documents.
- **Datasets from large repositories:** ELG, CLARIN, GitHub, HuggingFace, etc.
- **New datasets** compiled from public administrative and governmental data, websites and technical documentation, media websites, open science portal.

Procedures for improving the quality of the dataset:

★ **Deduplication:** prefiltering based on metadata, MinHash algorithm.
★ **PII identification:** returning number and ids of sentences containing sensitive information.
★ **Bias information:** returning number and ids of sentences containing potential abusive or biased content.

Source	# texts	# tokens	License
MARCELL	25K	45M	Public domain
CURLICAT	113K	35M	Creative Commons (CC)
BulNC Administrative	17K	79M	Public domain
BulNC Wikipedia	89K	41M	CC / GNU
BulNC Subtitles	146K	27M	OPUS

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

⊙ 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/