



# 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:

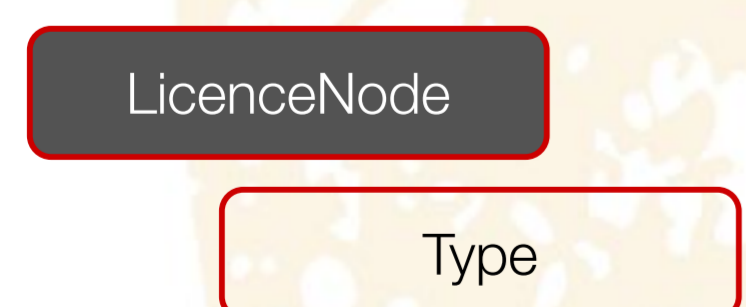
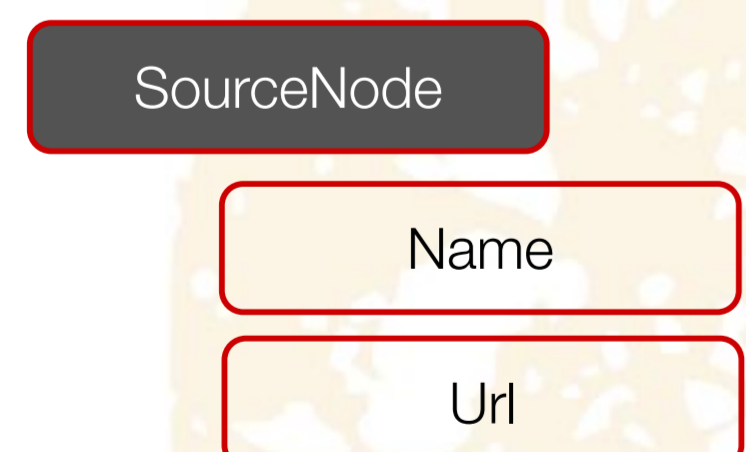
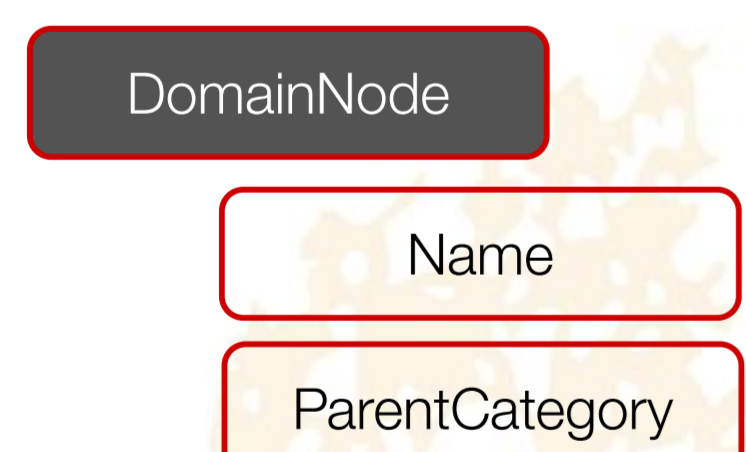
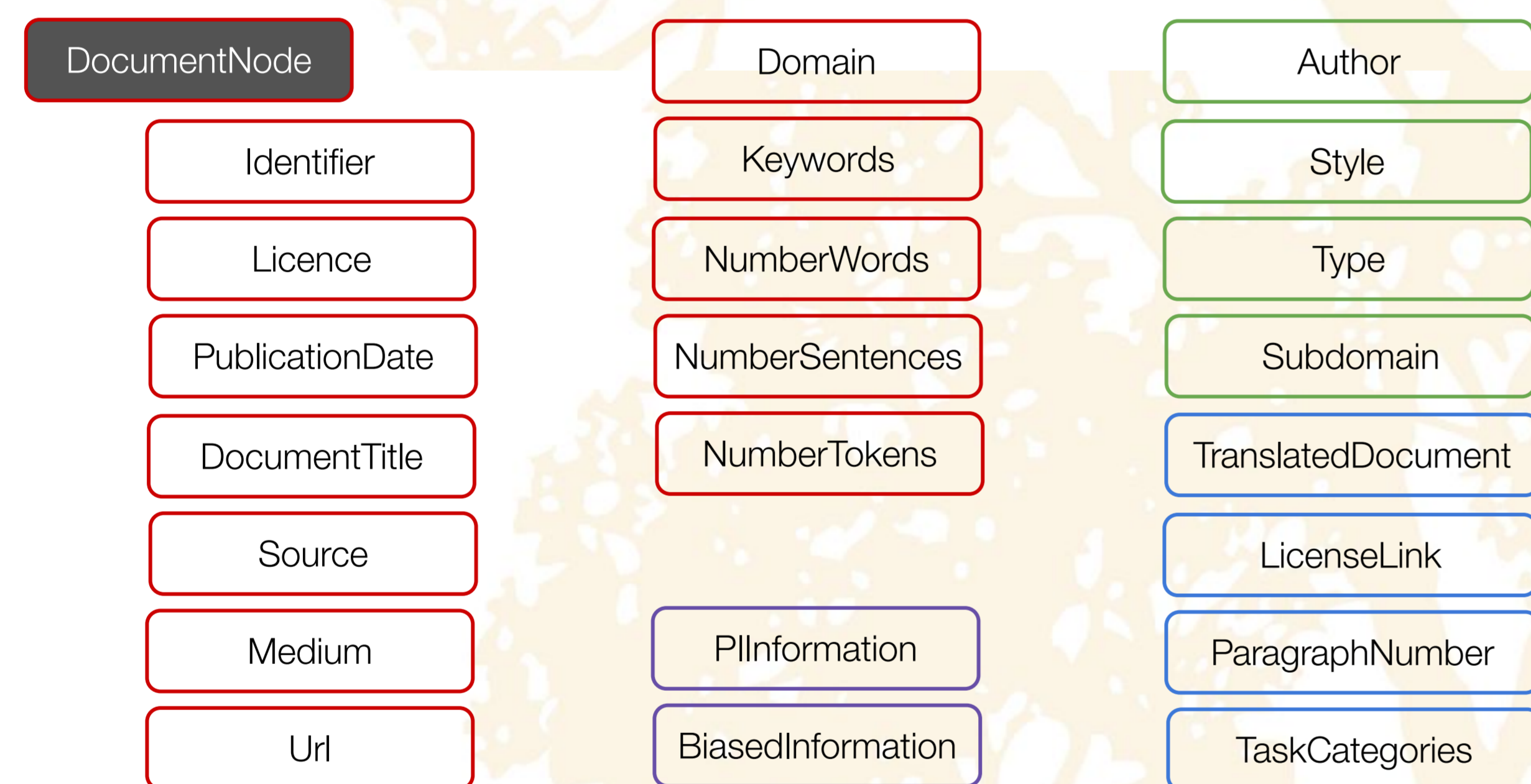
- **Data scarcity** – few sources for compiling large datasets for pre-training and fine-tuning LLMs.
- **Copyright restrictions** – limited datasets with no copyright issues, or legislation is unclear.
- **Quality of the data** – freely accessible data is often noisy and inhomogeneous.

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.



### Relations

Document – Domain	BELONGS_TO
Domain – Domain	SUBCATEGORY_OF
Document – Licence	LICENSED_WITH
Document – Author	WRITTEN_BY
Document – Source	PUBLISHED_IN

### IfGPT Dataset Search

#### License Types:

- Public Domain
- CC0
- CC-BY-SA
- CC-BY-NC-SA
- CC-BY
- CC-BY-NC
- CC-BY-ND
- GNU

#### Copyright Restricted

#### Domains:

- Economics
- Business
- School
- Healthcare
- Sociology
- Commerce
- Science
- Legislation
- Politics
- Education
- Leisure
- Subtitles
- Law
- Administration
- History
- Government

#### Year Range:

2015 2017

#### Keywords (separated by a comma):

Enter keywords separated by commas

Search

## 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

### Search Results

Number of text samples found: 4570

Number of sentences: 272,195

Number of tokens: 6,560,464

Download Results (json)

#### Selected Documents:

- bg\_MARCELL\_bg-100032** - Решение № 3008-ПВР от 12 януари 2016 г. относно утвърждаване образци на изборните книжа за произвеждане на избори за президент и вицепрезидент на Република България  
<http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100032>
- bg\_MARCELL\_bg-100062** - Постановление № 3 от 13 януари 2016 г. за създаване на Съвет по прилагане на Актуализираната стратегия за продължаване на реформата в съдебната система  
<http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100062>
- bg\_MARCELL\_bg-100067** - адм. д. № 14071/2015 г.  
<http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100067>
- bg\_MARCELL\_bg-100068** - адм. д. № 9243/2015 г.  
<http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100068>
- bg\_MARCELL\_bg-100069** - адм. д. № 13954/2015 г.  
<http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100069>

- 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/>

