IfGPT: A Dataset in Bulgarian for Large Language Models

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

Svetla Koeva[1], Ivelina Stoyanova[2], Jordan Kralev[3]
1,2: *Institute for Bulgarian Language, Bulgarian Academy of Sciences*
3: *IBL – Bulgarian Academy of Sciences / Technical University of Sofia*

svetla@dcl.bas.bg | iva@dcl.bas.bg | jkralev@dcl.bas.bg

LowResNLP 2025 | 13 September 2025

Motivation

Developing datasets for LLMs is a major challenge for languages with limited resources:

- **Data scarcity** – there are few sources for compiling large datasets for pre-training and fine-tuning LLMs;
- **Copyright restrictions** – difficult to find datasets that do not raise copyright issues;
- **Quality of the data** – freely accessible data is often noisy and inhomogeneous; procedures for data cleansing and selecting only high-quality texts further limit the scope of the data.

Objective

The **main objective** is to compile a large dataset **IfGPT** for Bulgarian combining existing corpora and datasets with newly compiled datasets, ensuring the texts are clean, deduplicated and of high-quality, supplied with extensive metadata.

The aim is to avoid redundant compilation of datasets by different users and multiple efforts required to cleanse the data and facilitate reusing the data to solve different application tasks.

Existing large datasets

There are many large and widely used text databases:

- **CommonCrawl** – raw web page data and metadata; massive quantity of data but low quality; derived from CommonCrawl are also:
  - **OSCAR (Open Super-large Crawled Aggregated coRpus)** – large multilingual corpus created by language classification and filtering of the CommonCrawl dataset.
  - **C4** and **mC4** – created using heuristic methods to filter out non-linguistic content and underwent extensive deduplication.
  - **CC100** – provides monolingual data for more than 100 languages.
- **Pile –** an 825 GB English text corpus developed for LLM training.
- **MassiveText** – a collection of large English language text datasets from various sources, including websites, books, news articles and code.

Existing large datasets

However:

- They rarely include Bulgarian data.
- The multilingual data (including Bulgarian) is a very small proportion of the dataset.
- Most of the existing datasets have already been included in datasets for LLM pretraining.

Existing datasets of Bulgarian

- **Bulgarian National Corpus (BulNC, 420 mln. tokens)** contains a wide range of texts of different sizes, different styles, time periods (synchronous and diachronic) and licences. Each text in the collection is labelled with metadata.
- **General News in Bulgarian (600 mln. tokens)** contains news from different thematic domains. The news items and their metadata were collected automatically from various (mainly Bulgarian) Internet sources, approx. 2 mln. web pages.

Existing datasets of Bulgarian

- **Bulgarian CURLICAT** – Curated Multilingual Language Resources for CEF.AT (**35 mln. tokens**) consists of texts from various sources divided into seven thematic domains: Culture, Education, European Union, Finance, Politics, Economy and Science.
- **Bulgarian MARCELL –** Multilingual resources for CEF.AT in the legal domain (**45 mln. tokens**) consists of legislative documents extracted from the Bulgarian State Gazette, documents from official institutions such as the government, the Bulgarian National Assembly, the Constitutional Court, etc.

Compiling new datasets of Bulgarian

There are additional sources for datasets in Bulgarian:

- ELG,
- CLARIN,
- GitHub,
- HuggingFace, etc.

Compilation of new datasets:

- Public administrative and governmental data,
- Websites and technical documentation,
- Media websites,
- Open science portal.

Improving the quality of IfGPT: Deduplication and cleaning up

Removing duplicate texts in the dataset improves the performance of LLMs. Two step procedure:

- Prefiltering based on metadata – year of publishing, source, etc.
- Main deduplication based on the MinHash and Locality Sensitive Hashing (LSH) algorithm.

Improving the quality of the texts:

- Removing boilerplate,
- Removing web elements (navigation, formatting, etc).
- Converting all into text format (pdf, documents, etc.). OCR texts avoided due to lower quality.

Improving the quality of IfGPT: PII and Bias information

Personally identifiable information is identified and handled as follows:

- MAPA anonymisation package for Bulgarian.
- Naive rule-based methods to detect sentences of the document with potentially sensitive information.
- The number of sentences with such information is counted and the output is the proportion of these sentences in the text.

Bias information is treated in a similar way:

- Potentially biased or abusive sentences are identified using lexical resources and rule-based methods.
- The number of sentences containing potential bias are counted and the output is the proportion of these sentences in the text.

IfGPT: Current structure

Current structure of the IfGPT dataset:

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

IfGPT: Metadata management

The metadata are stored in a **Neo4J** graph database with a schema capturing the key metadata entries and their connections.

- **Document nodes** with properties describing the source and properties of the text.
- **Author nodes** providing details of the authors, biography, etc.
- **Domain nodes** defining a shallow hierarchical structure of domains.
- **Licence nodes** defining the licence used.
- **Source nodes** providing the name and url of the source.

IfGPT: Metadata management

DocumentNode

- Identifier
- Licence
- PublicationDate
- DocumentTitle
- Source
- Medium
- Url

- Domain
- Keywords
- NumberWords
- NumberSentences
- NumberTokens

- PIInformation
- BiasedInformation

- Author
- Style
- Type
- Subdomain
- TranslatedDocument
- LicenseLink
- ParagraphNumber
- TaskCategories

IfGPT: Metadata management

DomainNode

- Name
- ParentCategory

SourceNode

- Name
- Url

LicenceNode

- Type

Relations

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

IfGPT: Online search interface

IfGPT Dataset Search

License Types:

- [x] Public Domain
- [x] CC0
- [ ] CC-BY-SA
- [ ] CC-BY-NC-SA
- [ ] CC-BY
- [ ] CC-BY-NC
- [ ] CC-BY-ND
- [ ] GNU
- [ ] Copyright Restricted

Domains:

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

Year Range:

2015 — 2017

Keywords (separated by a comma):

Enter keywords separated by commas

Search

IfGPT: Online search interface

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

bg_MARCELL_bg-100071 - *Решение за освобождаване от длъжност на инспектор в Инспектората към Висшия съдебен съвет*
http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100071

IfGPT: Online search interface

Search Results JSON

Generated on 9/11/2025, 8:36:15 AM


```
{
  "searchResults": {
    "statistics": {
      "numberOfTextSamples": 4570,
      "numberOfSentences": 272195,
      "numberOfTokens": 6560464
    },
    "selectedDocuments": [
      {
        "year": 2016,
        "PublicationDate": "2016-01-12",
        "Keywords": "",
        "NumberTokens": 118,
        "CollectionDate": "",
        "TranslatedDocument": "",
        "LicenceLink": "https://elrc-share.eu/static/metashare/licences/CC0.pdf",
        "TaskCategories": "",
        "Source": "",
        "Url": "http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100032",
        "BiasedInformation": "",
        "NumberWords": 107,
        "Type": "Решения на ЦИК",
        "Identifier": "bg_MARCELL_bg-100032",
        "Subdomain": "",
        "Medium": "text",
        "PersonallyIdentifiableInformation": "",
        "NumberParagraphs": 8,
        "Licence": "CC0",
        "Author": "Ministries and other institutions",
        "Style": "Legal",
        "DocumentTitle": "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",
        "Domain": "Government",
```

IfGPT: Future development

The future development of the **IfGPT** dataset includes:

- Expanding the dataset with new text samples.
- Completing the metadata description for some empty metadata categories.
- Detailed description of Task Categories for which given text sample is suitable (e.g. question answering).

IfGPT is created as a large dataset equipped with rich metadata for efficient search and retrieval of suitable documents, clearly defined tasks and thematic domains.
It enables fast and efficient fine-tuning of LLMs and RAG.

Acknowledgments

- 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/.

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

- https://ifgpt.dcl.bas.bg