Recent developments of the Bulgarian National Corpus

Svetla Koeva, Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
{svetla,iva}@dcl.bas.bg

Abstract

We present recent developments in the Bulgarian National Corpus, including data collection from various sources, cleaning of diverse datasets, enrichment with multimodal data, and extensive metadata, which resulted in the development of IfGPT, a large BulNC-based dataset. Typical methods for distributing the BulNC-based dataset are briefly described, with emphasis on effective searching within the metadata stored in a graph database.

1. Introduction

Over the past ten years, the Bulgarian National Corpus (BulNC) has undergone several developments to provide broader coverage of linguistic data and greater applicability to various NLP tasks. Since its establishment in 2009, the key features of BulNC have included **diversity of data** in terms of registers, domains, time periods, authors, and more; **multilinguality**; **extensive metadata** description; and **linguistic integrity**.

The significant development of BulNC, together with other national corpora in recent years, has been driven by the availability of large amounts of accessible data and new technologies for data collection, visualisation, and extraction of language facts and dependencies. Key areas of progress include the compilation and use of large volumes of multilingual and, to some extent, multimodal data; moving beyond simple corpus search and analysis to offer customised functions for linguistic analysis, such as defining words, tracking usage, detecting semantic shifts, and creating examples; serving as clean data for LLM pre-training and fine-tuning; and being analysed with LLMs.

The architectures of certain corpus management platforms enable the simultaneous processing of texts containing billions of words in many languages. For example, Sketch Engine provides access to over eight hundred corpora in more than one hundred languages, and allows complex linguistic queries and services (Kilgarriff et al., 2014).[1] English-Corpora.org is a collection of large corpora of English and its varieties, several of which contain billions of words (Davies, 2025).[2] The Czech National Corpus[3] provides access to written, spoken, parallel, and diachronic corpora comprising several billion words, which can be queried via the KonText interface (Machálek, 2020). The German Reference Corpus DeReKo, the largest corpus of written German, contains more than 60 billion words,[4] and is accessible through the KorAP corpus analysis platform (Diewald et al., 2016).

Recently, LLMs have been integrated into corpus query tools such as AntConc (Anthony, 2024), allowing identification of missing information and suggesting corrections for inconsistencies in dictionary drafts.

In addition to providing public access to the BulNC data for linguistic research, over the past ten years our efforts have focused on expanding BulNC with diverse and linguistically clean data suitable for NLP research and, more recently, for pre-training and fine-tuning LLMs. This has resulted in a shift in dataset accessibility, enabling the extraction of subcorpora for specific tasks based on extensive metadata.

These efforts have led to the development of the large **BulNC-based dataset** within the project *IfGPT: Infrastructure for Fine-tuning Pre-trained Large Language Models*[5] (also called the **IfGPT dataset**), with a special focus on the efficient management of large text data.

Alongside the expansion of textual data, we aim to provide more diverse data in terms of multilinguality (parallel corpora), various levels of annotation (including aligned corpora), and multimodal corpora suitable for a wide range of NLP and AI applications.

2. IfGPT, a large BulNC-based dataset

The components of **IfGPT, a BulNC-based dataset**, can be categorised into three main groups according to text type, composition, and potential uses: 1) collections of texts that have already been created, processed, and are available (BulNC belongs here); 2) other existing datasets of Bulgarian

[1] https://www.sketchengine.eu/
[2] https://www.english-corpora.org/
[3] https://www.korpus.cz/
[4] https://korap.ids-mannheim.de/
[5] https://ifgpt.dcl.bas.bg/en/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 71–75. Palma, Mallorca. May 11, 2026.

texts that need to be reviewed, downloaded, and, if necessary, have their text and metadata formats converted to those of the IfGPT dataset; 3) compilation of new datasets through targeted crawling and processing of identified texts for filtering, cleaning, deduplication, and addition of metadata.

The new additions to IfGPT were collected, cleaned, and processed mainly within projects funded at national or European level, such as:

- **CEF Automated Translation for the EU Council Presidency**,[6] which involved collecting a large amount of parallel data and terminological resources in Bulgarian and English to train machine translation systems, focusing on official communication and translation challenges during the EU presidency.
- **Multilingual Resources for CEF.AT in the Legal Domain** (**MARCELL**),[7] which collected national legislative texts in seven languages – Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian – and annotated each subcorpus with morphosyntax, dependency structure, named entities, and IATE/EuroVoc terminology.
- **Curated Multilingual Language Resources for CEF AT** (**CURLICAT**),[8] which collected, cleaned, anonymised, and annotated licence-free texts in the same seven languages, producing over 14 million sentences across the health, culture, science, and government domains, with a harmonised metadata schema designed for neural machine translation.
- **Infrastructure for Fine-tuning Pre-trained Large Language Models** (**IfGPT**), which aims, among other tasks, to collect, filter, anonymise, and deduplicate large, diverse, high-quality text data for fine-tuning pre-trained LLMs for Bulgarian.

Further extensions of the dataset include newly collected and processed texts from various time periods. Older texts, such as news articles, periodicals, and books published before 1990, are also collected and processed using OCR. Table 1 shows the most important parts of the IfGPT dataset.

Source	# texts	# tokens	Licence
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	289M	various
EN internet	45K	8,144M	various
News up to 1990	5,544K	270,52M	various
Periodicals up to 1990	25K	30M	various
New periodicals	4,119K	4,378M	various
Books	22K	630M	various

Table 1: Current structure (March 2026). Licences: PD – public domain, CC – Creative Commons (various), GNU – GNU free license, various – other (including restrictive) licences.

3. Multimodal Data

The BulNC-based dataset is extended with multimodal data from the **Multilingual Image Corpus (MIC21)** (Koeva et al., 2022). MIC21 provides pixel-level annotations for over 203,000 objects in more than 21,000 images, covering 730 object classes across four thematic domains.[9]

The images are carefully selected to ensure high-quality, copyright-free content from thematically related domains (Sport, Transport, Art, and Security), comprising 130 related subdomains, and are supplied with available metadata. Annotation is performed by drawing or correcting automatically generated polygons using the Detectron2 model (Wu et al., 2019)), from which bounding boxes are then generated automatically. This enables wide application of the dataset in various computer vision tasks: image classification, recognition and classification of single objects in an image, or of all object instances in an image (semantic segmentation). An example of an image from the domain **Art** and the subdomain **Violinist**, with three annotated objects (*violinist*, *violin*, and *bow*), is shown in Figure 1.

The classes for object annotation are organised in an Ontology of Visual Objects (Koeva, 2022), which offers options for extracting relationships between objects in images, constructing diverse datasets with varying levels of object class granularity, and compiling suitable sets of images illustrating different thematic domains. Some classes and relations are inherited from WordNet. Additional classes and relations are included in the ontology if they are not present in WordNet; for example, ***Bowler*** *wears* ***Bowling shoes***. Object classes are linked to certain metadata values of the BulNC-based dataset, thereby relating visual content to the textual dataset.

The object labels from the Ontology are linked to their synonyms, definitions, and usage examples in 25 languages. The selection of languages was

[6] https://tilde.ai/machine-translation/
[7] https://marcell-project.eu
[8] https://curlicat.eu
[9] https://dcl.bas.bg/en/projects_list/mic21/

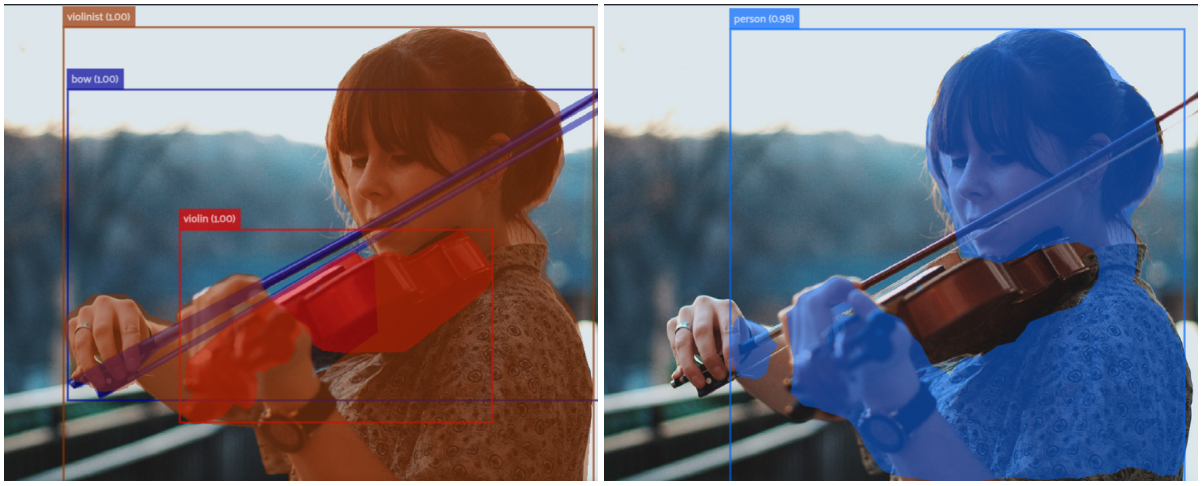

Figure 1: Image with masks, bounding boxes and labels: on the left – MIC21 manually annotated; on the right – automatically annotated using the Detectron2 model.

based on the availability of wordnets in various languages in the Extended Open Multilingual Wordnet (Bond and Foster, 2013). Where WordNet translations are unavailable, additional sources of translations are used, mainly BabelNet[10] and machine translation. The labels of objects whose concepts are not present in WordNet have been translated by experts only into English and Bulgarian. The multilingual layer makes the dataset suitable for artificial intelligence applications such as multilingual image captioning, question answering, and machine translation of multimodal content.

Recently, images have been provided with short narrative descriptions that explain the relationships between the depicted objects.

4. IfGPT dataset processing pipeline

For the expansion of the BulNC, the integration of resources developed through collaborative international projects, and the preparation of data tailored for language technologies and LLMs, the following components of the **IfGPT dataset processing pipeline** have been developed:

- **File handling module** for managing files in appropriate formats for text and metadata (e.g., plain text, JSONL, CSV), using the adopted metadata schema.
- **Dataset quality maintenance module** providing functions for string manipulation, data cleaning, and error handling to support data quality assurance through text deduplication, identification and labelling of personally identifiable information, and detection of potential bias.
- **Metadata extraction module** for obtaining metadata from the document source and content, and providing appropriate metadata descriptions.
- **Annotation module** for introducing traditional linguistic annotation in CoNLL-U Plus format (Koeva et al., 2020).
- **Dataset construction module** for creating subdatasets for specific purposes based on extensive metadata.
- **Search module** providing an online interface for browsing metadata values for selection (Koeva et al., 2025). Its output is either a newly constructed subdataset or a selection of links for downloading relevant parts of the subdataset.

5. Metadata and ways of distribution

The metadata is designed for searching and retrieving information to support various research and applications, and therefore has a complex graph-based structure of related categories (Koeva et al., 2016). The metadata of the IfGPT dataset originates from the BulNC and is harmonised with the metadata of newer multilingual corpora such as MARCELL and CURLICAT.

The metadata includes 15 mandatory categories covering technical details (such as identifier), source description (source URL, licence), and document statistics (number of words, sentences), as well as 9 optional categories describing features of the document (such as author, style). The metadata is managed using the graph database Neo4J,[11] which is designed to handle large vol-

[10] https://babelnet.org/

[11] https://neo4j.com/

umes of interconnected data efficiently and maintains performance under complex queries using the Cypher query language (Francis et al., 2018). The graph database effectively models relations between metadata values (e.g. `WRITTEN_BY` for authorship, `LICENCED_WITH` for licensing, `BELONGS_TO` for domain classification) and allows efficient access to the metadata and extraction of different subsets according to users' needs.

The Bulgarian National Corpus offers a customised web interface for searching the corpus, building concordances, and extracting examples (Koeva et al., 2012, 100-101).[12] The search system supports complex linguistic queries involving different levels of annotation (POS, morphosyntactic features, semantic relations) combined in various ways.

Parts of the BulNC and the extended IfGPT dataset with open licences are available for direct download, while some parts are subject to copyright restrictions. The latter may be used to compile subsets for specific users and tasks, but cannot be redistributed directly.

The IfGPT Dataset search interface[13] allows users to browse and filter the large collection of clean, deduplicated Bulgarian text documents by several criteria: type of licence, domain, time period, and keywords. The user can export metadata, links, or raw texts. It is aimed for dataset compilation and extraction of data for use in NLP applications and LLM fine-tuning.

6. Conclusion and future work

In summary, the IfGPT dataset, based on BulNC, is designed to be as large as possible while incorporating rich metadata to support efficient search and retrieval of relevant data for research and practical applications, including the pre-training and fine-tuning of large language models. Future development of IfGPT will include improvements in data distribution and curation methods, such as integrating large language models to identify missing information, resolve inconsistencies, and enhance the overall dataset compilation process.

Extremely large text collections have been developed by crawling internet data. For example, Common Crawl contains petabytes of data, including Bulgarian. It includes raw web page data, metadata extracts, and text extracts (Common Crawl Foundation, 2025). Many LLM datasets have been created based on it, such as mC4, OSCAR, and CulturaX. One of the largest, CulturaX, is a multilingual dataset with 6.3 trillion tokens in 167 languages, including Bulgarian. The dataset undergoes extensive cleaning and deduplication. The HPLT Corpus (Burchell et al., 2025) contains monolingual corpora covering 193 languages, including Bulgarian, and approximately 8 trillion tokens. Parallel corpora from monolingual data for 50 languages paired with English were derived, containing over 380 million sentence pairs. The corpus was extracted from 4.5 petabytes of Internet Archive and Common Crawl data.

Ensuring the quality of large datasets is a critical prerequisite for efficient and reliable training of LLMs, as noise, redundancy, and malformed content directly degrade model performance and introduce systematic biases (Kreutzer et al., 2022). For Bulgarian, a morphologically rich and low-resource language, ensuring high data quality is challenging due to limited tools for efficient identification of text matches and near matches, inconsistent orthography, incomplete sentences, and texts produced using machine translation. Deduplication is especially important at scale, particularly for near-duplicate documents, which can artificially increase corpus size and cause language models to over-represent certain linguistic patterns.

Developing robust automatic methods for these tasks for Bulgarian is difficult, as pipelines implemented for high-resource languages are not always suitable or straightforward to adapt, while building language-specific solutions requires significant resources, including manual annotation and evaluation. Thus, developing efficient and scalable methods for quality checks and improvement of large datasets in Bulgarian remains an open research challenge.

7. Acknowledgments

The present study is carried out within the project Infrastructure for Fine-tuning Pretrained Large Language Models, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Laurence Anthony. 2024. Antconc (version 4.3.0).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komu-

[12] https://search.dcl.bas.bg/

[13] https://ifgpt.dcl.bas.bg/ifgpt-dataset/

lainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Common Crawl Foundation. 2025. Common crawl web corpus.

Mark Davies. 2025. English-corpora.org.

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP architecture – diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3586–3591, Portořož, Slovenia. ELRA.

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1433–1445, New York, NY, USA. Association for Computing Machinery.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1(1):7–36.

Svetla Koeva. 2022. Ontology of visual objects. In *Proceedings of the Fifth International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pages 120–129, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2022. Multilingual image corpus – towards a multimodal and multilingual dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. ELRA.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2025. IfGPT: A dataset in Bulgarian for large language models. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 65–75, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 1(1):65–110.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva, and Tsvetana Dimitrova. 2016. Metadata extraction, representation and management within the Bulgarian National Corpus. In *Priceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, pages 33–39. ELDA.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 7003–7008, Marseille, France. ELRA.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. GitHub.