IfGPT, a large dataset representing Bulgarian, with the Bulgarian National Corpus as its core

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

Abstract

The paper introduces the IfGPT dataset, which integrates several Bulgarian text collections, including the Bulgarian National Corpus, and applies cleaning, deduplication, and LLM-oriented metadata such as personally identifiable information and bias scores. The composition of the IfGPT dataset is presented, along with the unified metadata schema and metadata management in a graph database, enabling efficient querying and document selection for specific tasks. The main contributions are the integration of multiple Bulgarian text collections into a unified dataset, the development of a standardised metadata schema with graph-based organisation, and the provision of efficient metadata querying mechanisms to support LLM development.

1. Introduction

The development and management of large-scale corpora present interconnected technical, linguistic, and management challenges. From a technical perspective, scalable storage and efficient retrieval of text, metadata, and annotation layers are essential (Mohammadi et al., 2025). From a linguistic and NLP perspective, large reference corpora must be diverse and represent a wide range of language use, including low-resource languages, under-represented phenomena, and historical texts.

The **Bulgarian National Corpus** (**BulNC**)[1] is a standard reference corpus designed to reflect the natural distribution of the Bulgarian language across text types, genres, styles, and time periods, ensuring domain coverage and distributional balance (Koeva et al., 2012). Its main purpose is to support linguistic studies focused on the lexical and grammatical features of Bulgarian, dictionary creation, and the exploration of language change. The Bulgarian texts are annotated using the Bulgarian natural language processing pipeline (Koeva et al., 2020), which integrates several tools for different layers of annotation: tokenisation, part-of-speech tagging, lemmatisation, dependency parsing, word sense annotation, lexical relations (synonyms, hypernyms, and similar adjectives), noun phrase identification, and named entity recognition. Linguistic integrity is maintained by deduplicating texts and removing documents with typographical errors, incomplete sentences, or malformed words.

The large volume of available data, including for Bulgarian, has enabled the development of datasets that encompass linguistic and human knowledge about the world for training large language models (LLMs). The dominant approach is to collect as much data as possible, mainly from the web, and then filter it by cleaning and deduplication, as well as by removing content that could degrade model behaviour, such as toxic content, personally identifiable information (PII), near-duplicate documents that may cause memorisation, and very low-quality text that may introduce noise.

The efforts to develop, maintain, expand, and improve the Bulgarian National Corpus are naturally combined with the compilation, cleaning, maintenance, and enhancement of large Bulgarian datasets for pre-training and fine-tuning LLMs. These efforts have resulted in the creation of the large **BulNC-based dataset** within the project **IfGPT: Infrastructure for Fine-tuning Pre-trained Large Language Models**[2] (the **IfGPT dataset**).

The large IfGPT dataset integrates several Bulgarian datasets, including the Bulgarian National Corpus. Like the BulNC, IfGPT contains authentic Bulgarian language data that is cleaned and deduplicated. The IfGPT description adds some LLM-oriented metadata (i.e. PII scores, bias scores) that a pure reference corpus such as BulNC does not include. Annotation in the IfGPT dataset is also modest compared to the BulNC, limited to sentence markup, which relates to the main purpose of its intended use.

In the next section, we briefly present related work on the development of reference corpora and the position of the BulNC within this context. The main part of the paper is devoted to the composition of the IfGPT dataset, its metadata description, and its management through a graph database that provides various access options.

The main contributions of this work are as fol-

[1] https://dcl.bas.bg/bulnc/

[2] https://ifgpt.dcl.bas.bg/en/

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 12–24. Palma, Mallorca. May 11, 2026.

lows. First, we merge several large Bulgarian text collections into a unified dataset with standardised metadata and text formats. Second, we provide a unified metadata schema for all documents and organise the metadata categories in a graph-based representation. Finally, we offer efficient mechanisms for querying metadata to identify suitable documents for specific tasks such as LLM fine-tuning or Retrieval-Augmented Generation (RAG).

2. Bulgarian National Corpus and related work

The rapid growth of large-scale language data in recent years has advanced in several directions: expanding existing reference corpora with new text types, integrating multilingual and multimodal data, and producing training data specifically tailored for language technologies and large language models.

2.1. Corpus Query Tools

Most national corpora have an online search interface linked to a predefined document set. For some corpora, a dedicated corpus query tool has been developed, such as the PELCRA search engine for the National Corpus of Polish (Pęzik et al., 2016) and the KorAP corpus analysis platform (Diewald et al., 2016) for the German Reference Corpus DeReKo, among others.

Other corpora use open-source corpus query tools: KonText (Machálek, 2020) is a web-based corpus query tool for working with texts in the Czech National Corpus. The Slovak National Corpus is accessed via the NoSketch Engine (Rychlý, 2007; Kilgarriff et al., 2014). The Corpus of Written Standard Slovene is available via Sketch Engine, NoSketch Engine, and KonText (Krek et al., 2020).

In many corpus query tools, users can select which corpus to use and create their own collections of texts within the corpora by filtering according to various criteria, such as topic, subgenre, author, or source, and then search this subcorpus as if it were their own corpus. Tools such as Sketch Engine, NoSketch Engine, KonText, and AntConc (Anthony, 2024) allow access to multiple corpora, either as a single option or as a selection of subcorpora.

The Bulgarian National Corpus has a web interface for searching the corpus,[3] building concordances, and extracting examples (Koeva et al., 2012, 100-101). The search system allows complex linguistic queries involving different levels of annotation combined in various ways. It was designed to support both monolingual and parallel corpora in a uniform way. Compared to CQL, the implemented Designed Query Language (DQL) supports terms such as word, relation (i.e. word form, synonym, hypernym, etc.), and their combinations. Both ordered and unordered queries are supported, as well as conjunction and disjunction of ordered queries, such as searching for paraphrases.

[3] http://search.dcl.bas.bg

2.2. Expanding large reference corpora

Large reference corpora have increased in both volume and the range of text types they represent. Egbert et al. (2022) provide a systematic methodological overview of large corpora, arguing that size alone does not guarantee representativeness and proposing a two-pillar framework: domain coverage and distributional balance. The authors criticise the assumption that "bigger is always better", redirecting efforts from scale to design.

Hashimoto and Nelson (2024) examine 709 corpus descriptions published between 2010 and 2019. The authors analyse sampling decisions and the methodological principles employed by recent large-scale expansions, which rely on growth in corpus size accompanied by transparent, principled sampling techniques (Egbert et al., 2022; Hashimoto and Nelson, 2024).

Over the past 10 years, the main effort in developing the Bulgarian National Corpus through various national and international projects has focused on collecting, cleaning, and enriching large datasets, which will be discussed in more detail in the following section.

2.2.1. Large collections of unstructured data

The dominant paradigm for LLM pre-training data is large-scale web harvesting, with Common Crawl serving as the primary raw source for most major datasets. Conneau et al. (2020) introduced the CC-100 dataset – approximately two terabytes of filtered monolingual text in one hundred languages derived via the CCNet pipeline. Subsequent efforts have focused on aggressive quality filtering and deduplication.

Gao et al. (2020) established the multi-source paradigm with The Pile – 825 GiB of English text from 22 curated subsets spanning books, code, scientific papers, and online discussion – showing that domain diversity substantially improves downstream generalisation.

Soldaini et al. (2024) released Dolma (3 trillion tokens across six source types), providing both the data and the complete processing toolkit. Nguyen et al. (2024) released CulturaX (6.3 trillion tokens in 167 languages) by merging and cleaning mC4 and OSCAR. Singh et al. (2024) produced the Aya Collection, aggregating 513 million instances across 114 languages through an open participatory science model.

The Bulgarian National Corpus was expanded with several domain-specific corpora from the

OPUS collection (Tiedemann, 2012). The largest of these are the EMEA corpus of administrative medical texts and the OpenSubtitles corpus (film subtitles) (Lison and Tiedemann, 2016).

2.2.2. Special purpose datasets

Alongside general-purpose pre-training corpora, an increasing body of work focuses on developing datasets for specific task types or application contexts.

Instruction tuning has become a particularly active area for special-purpose dataset construction. Chung et al. (2022) introduced the FLAN v2 collection, comprising more than 1,800 reformatted tasks. Zhou et al. (2023) developed LIMA, a set of 1,000 carefully hand-selected prompt–response pairs that produce a competitive instruction-following model. Taori et al. (2023) operationalised the self-instruct paradigm by generating 52,000 instruction examples from GPT-3 and releasing both the data and training code. For multilingual instruction data, Muennighoff et al. (2023) present the ROOTS corpus (1.6 TB of text across 46 natural and 13 programming languages), which underpins the BLOOM model. Similarly task-focused, Kocoń et al. (2025) document the CLARIN-PL infrastructure, including the MultiEmo sentiment corpus extended across eleven languages.

Together, these efforts mark a shift from passive data collection to active dataset design oriented towards particular tasks, ensuring prompt diversity and quality control.

Within the structure of the BulNC, the Diachronic Corpus of Bulgarian has been compiled to support research on the lexical and grammatical features of Bulgarian over time.[4]

The corpus contains texts totalling 1.1 million words from 1851 to recent years, divided into six time intervals (1851–1880; 1881–1910; 1911–1930; 1931–1950; 1951–1990; 1991–2021), and covers three domains: fiction, news, and science. The texts are sourced from various places, including scanned copies of periodicals, international databases (e.g. Gutenberg), and modern electronic databases for texts from 1990 onwards. The choice of domains was based on observations of domain coverage across time periods. Administrative and other types of texts are rare in the earlier periods and are therefore not included in the Diachronic Corpus.

2.2.3. Multilingual Large Datasets

The development of large-scale multilingual corpora has involved sharing annotation schemes, frameworks, and data processing pipelines.

The TenTen corpus family (Jakubíček et al., 2013) covers more than fifty languages, each with over ten billion words, and has progressively added genre and topic classification in its latest releases.

The ParlaMint project (Erjavec et al., 2024) uses a shared Parla-CLARIN scheme, Universal Dependencies annotation, and named-entity labels. This results in a comparable corpus in which political discourse can be studied across languages and political systems.

Hundreds of parallel text corpora are already available with Bulgarian as one of the languages, most of which can be downloaded from repositories such as ELG[5] and CLARIN.[6] The bilingual corpora mainly contain Bulgarian and English or other European language pairs, for example, Bulgarian – Modern Greek, Bulgarian – German, Bulgarian – French, Bulgarian – Italian, and Bulgarian – Spanish.

Bulgarian is included in even more multilingual corpora, some of which are sentence-aligned, enabling straightforward cross-lingual research. Many large multilingual corpora are created automatically from web sources (e.g. Common Crawl, Wikipedia), while others are compiled from institutional, parliamentary, subtitle, or legislative data.

3. IfGPT composition

The IfGPT dataset is a collection of datasets, primarily in Bulgarian but also in English, based on the Bulgarian National Corpus. As the structure of BulNC has been described in detail elsewhere (Koeva et al., 2012), we focus here only on the main components that currently comprise the IfGPT dataset. These are summarised in Table 1.

The dataset **Bulgarian MARCELL** consists of legislative documents divided into fifteen types (Váradi et al., 2020). The documents span from 1946 to 2023 and were extracted from the Bulgarian State Gazette, the official gazette of the Bulgarian government, which publishes documents from official institutions such as the government, the Bulgarian National Assembly, the Constitutional Court, and others. The Bulgarian dataset contains 25,283 documents categorised into eleven types: Administrative Court; Agreements; Amendments (legal acts); Conventions; Decrees; Decrees of the Council of Ministers; Directives; Instructions; Laws (legal acts); Memoranda; Resolutions. The dataset comprises approximately 45 million tokens and 3,281,000 sentences (as of the end of March 2021). Bulgarian MARCELL is part of a comparable corpus of national legislative documents for seven languages (Bulgarian, Croatian, Hungarian, Polish,

[4] https://dcl.bas.bg/bulnc/en/dostap/izteglyane/

[5] https://live.european-language-grid.eu/

[6] https://www.clarin.eu/

Dataset & Language(s)	Domains	Size	Format & annot.	Source & Licence
Bulgarian MARCELL BG 1946–2023	Legal (11 types: admin. court, agreements, amendments, conventions, etc.)	25K texts; 3.28M sents; 45M tokens	CoNLL-U+; morph., dep., NER, EuroVoc/IATE annotation	Bulgarian State Gazette Public Domain
Bulgarian CURLICAT BG	7 domains: Culture, Education, EU, Finance, Politics, Economics, Science	6K texts; 22.8M tokens	CoNLL-U+; JSON; full ling. annotation	BulNC; science sources: books, PhD theses; web CC-BY CC-BY-SA CC-BY-NC
Aligned and Normalised Parallel Data BG–EN	16 domains: General News, BG Presidency, Economics, Culture, Military, Politics, etc.	1.1M sent. pairs; 19.0M words (BG); 19.2M words (EN)	Sentence-aligned pairs; partly manual selection & correction	Web media; institutional websites Public Domain Various
General News in Bulgarian BG	185 domains	2.1M texts; 33.4M sents; 601M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (11.8K domains, 2.1M pages) Various
General News in English EN	185 domains	5.9M texts; 166.7M sents; 3.3B words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (324.5K domains, 5.9M pages) Various
News about the the Bulgarian Presidency in Bulgarian BG	185 domains	36.8K texts; 698K sents; 16.6M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (613 domains, 36.8K pages) Various
News about the Bulgarian Presidency in English EN	185 domains	12.3K texts; 292K sents; 8.8M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (663 domains, 12.3K pages) Various
General News in English from Bulgaria EN	General news (Bulgarian electronic media)	19.1K texts; 876K sents; 18.6M words	Sentence splitting; tokenisation; language detection; normalisation & cleaning	Web crawling (140 BG domains, 19.1K pages); predefined BG domain list Various
Filtered General News in English from Bulgaria EN	185 domains	19.1K texts; 237K sents; 5.5M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (140 BG domains, 19.1K pages); predefined BG domain list Various
Filtered News about the Presidency in English from Bulgaria EN	185 domains	1.4K texts; 20.6K sents; 504.6K words	JSON; metadata; automatic categorisation; normalisation & cleaning	Focused web crawling (55 BG domains, 1.4K pages) Various
Collection of Bulgarian Texts BG	4 styles; 42 domains: Adventure, Archaeology, Chemistry, Computers, Court, Culture, Ecology, Economics, etc.	66 collection files; 28.9M sents	Sentence splitting (bgLPC); no metadata; arbitrary sentence order	Internet Various
Collection of English Texts EN	4 styles; 13 domains: Court, Culture, Ecology, Economics, Health, History, etc.	45 collection files; 8.1M sents	Sentence splitting (bgLPC); no metadata; arbitrary sentence order	Internet Various
IfGPT – News BG Until 1990	News (historical)	5.5M texts; 270.5M tokens	OCR; LLM-assisted article separation & metadata extraction	Printed periodicals Various
IfGPT – Periodicals BG Until 1990	Periodicals (historical)	25K texts; 30M tokens	OCR and pagination and metadata extraction	Printed periodicals Various
IfGPT – New periodicals BG After 1990	Contemporary periodicals	4.1M texts; 4.4B tokens	Text & metadata extraction; LLM-assisted post-processing	Contemporary press Various
IfGPT – Books BG	General (books)	22K texts; 630M tokens	OCR and pagination; title-page metadata extraction	Printed books Various
Multilingual Image Corpus (MIC21) Image BG Various	4 domains (Sport, Transport, Arts, Security); 130 subdomains	22K images; 230M objects	annotated objects in images; short narrative descriptions	Open repositories CC-BY-SA

Table 1: Overview of the components in the IfGPT dataset with their key features. The Bulgarian language processing pipeline, is available at: http://dcl.bas.bg/dclservices/
.

Romanian, Slovak, and Slovenian) collected within the project **Multilingual Resources for CEF.AT in the Legal Domain** (**MARCELL**).[7] The Bulgarian MARCELL dataset is annotated in CoNLL-U Plus format (Koeva et al., 2020) for morphosyntax, dependency structure, and named entities. Documents are classified into thematic domains and enriched with specialised terminology identified from IATE[8] and EuroVoc.[9]

The dataset **Bulgarian CURLICAT** contains 113,087 documents divided into seven thematic domains: Culture, Education, European Union, Finance, Politics, Economics, and Science (Váradi et al., 2022a). The dataset comprises 6,036 documents with a total of 22,809,225 tokens. All documents are licensed under CC-BY, CC-BY-SA, or CC-BY-NC. To ensure a sufficient number of copyright-free documents, several sources were identified, including a library of scientific texts (books and PhD theses) and other websites providing texts from the required thematic domains. The texts are categorised, linguistically annotated, and provided in CoNLL-U Plus format, in the same way as the Bulgarian MARCELL dataset. The dataset was created within the CURLICAT project (**Curated Multilingual Language Resources for CEF.AT**) (Váradi et al., 2022b,a), which extended the approach to provide comparable corpora in domain-specific areas for the same seven languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian.[10]

Within the project **CEF Automated Translation for the EU Council Presidency**,[11] a dataset of **Aligned and Normalised Parallel Data** has been collected, curated and annotated, comprising the thematic domains presented in Table 2. These include parallel texts in English and Bulgarian aligned at sentence level. The parallel texts have been selected and collected from electronic media: Bulgarian National Radio, Bulgarian National Television, Bulgarian News Agency, Focus Information Agency, Sofia News Agency, web publications, web newspapers, and institutional websites with open access to relevant texts in Bulgarian and English.

The dataset **General News in Bulgarian** contains news from various thematic domains. The news and metadata were automatically acquired from different, predominantly Bulgarian, internet sources: 11,840 web domains and 2,116,739 web pages. The total number of words in the collected General News in Bulgarian is 601,330,975, distributed across 33,375,366 sentences. A crawling platform was used to identify and acquire monolingual data from websites, detect near-duplication at the document level, and normalise and clean the text. The resulting texts were structured into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 thematic domains, ordered by probability. The distribution of texts across the most frequent thematic domains is shown in Table3.

Domain	Sentence pairs	Words (BG)	Words (EN)
General News	3,118	61.9K	64.6K
BG Presidency	3,000	69.3K	74.6K
Science	454	9.2K	9.5K
Fiction	1,177	12.6K	13.3K
Economics	10.1K	201.6K	199.9K
Culture	3,164	54.6K	56.4K
Military	12.7K	265.8K	255.5K
Politics	66.2K	1.3M	1.3M
Social Domain	578	11.6K	11.6K
Undetermined	224.9K	4.6M	4.5M
Ecology	238.7K	4.9M	4.8M
History	5,215	98.6K	108.1K
Philosophy	725	15.3K	18.2K
Psychology	852	20.6K	25.2K
Medicine	511.4K	6.7M	7.0M
Technology	50.2K	712.6K	703.2K
Total	**1.1M**	**19.0M**	**19.2M**

Table 2: Overview of parallel texts in the **Aligned and Normalised Parallel Data** by domain

Domain	Number of texts
Economy	919,599
Sociology	735,566
Politics	718,540
Law	711,131
Enterprise	591,679
Commerce	455,647
Pedagogy	447,613
Administration	410,567
School	397,465
Free Time	365,550
History	356,974

Table 3: Distribution of texts in the most frequent domains in the **General News in Bulgarian** dataset.

The dataset **General News in English** contains news from various thematic domains. The news articles, together with some metadata, were automatically collected from numerous internet sources: 324,493 web domains and 5,961,124 web pages. The total number of words in the collected General News in English is 3,324,746,119, distributed across 166,718,125 sentences. A crawling platform was used to identify and acquire monolingual data from websites, detect near-duplication at the document level, and normalise and clean the text. The resulting texts were structured into JSON files,

[7] https://marcell-project.eu
[8] https://iate.europa.eu
[9] https://eur-lex.europa.eu
[10] https://curlicat.eu
[11] https://tilde.ai/machine-translation/

which contain extracted metadata and automatic categorisation of the content into 185 thematic domains (ordered by probability). The distribution of texts across the most frequent thematic domains is shown in 4.

Domain	Number of texts
Economy	4,516,499
Enterprise	3,895,116
Commerce	3,797,299
Exchange	2,879,980
Law	2,764,209
Finance	2,592,014
Bookkeeping	1,720,943
Banking	1,632,725
Politics	1,608,819
Sociology	1,597,035

Table 4: Distribution of texts in the most frequent domains in the **General News in English** dataset.

The dataset **News in Bulgarian about the Bulgarian Presidency of the Council of Europe** contains news thematically related to the Bulgarian Presidency. The news articles, together with some metadata, were automatically collected from various sources: 613 web domains and 36,835 web pages. The total number of words in the collected Bulgarian news is 16,550,562, distributed across 698,434 sentences. A crawling platform was used to acquire monolingual data from websites, as well as for normalisation, cleaning, and near-duplicate removal at the document level. The texts were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. The thematic domains most frequently assigned to the largest number of JSON files are: politics (27,089 files), economy (25,688 files), sociology (24,941 files), law (20,656 files), enterprise (19,526 files), administration (19,358 files), history (15,052 files), and diplomacy (11,167 files).

The dataset **News in English about the Bulgarian Presidency of the Council of Europe** contains news thematically related to the Bulgarian Presidency. The news articles, along with some metadata, were automatically acquired from various sources: 663 domains and 12,327 pages. The total number of words in the collected news in Bulgarian is 8,794,285, distributed across 292,111 sentences. A crawling platform was used to acquire monolingual data from websites, as well as for normalisation, cleaning, and near-duplication detection at the document level. The texts were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. The domains most frequently assigned to the largest number of JSON files are: economy (5,590 files), politics (5,531 files), enterprise (4,600 files), law (4,488 files), sociology (4,484 files), administration (4,011 files), diplomacy (3,700 files), finance (3,464 files), history (2,953 files), commerce (2,672 files), and exchange (2,287 files).

The following three datasets are sourced specifically from the Bulgarian web domain. Filtering between Bulgarian and non-Bulgarian web domains was conducted by extracting the country code from the WHOIS (BG) database, identifying the IP GeoLocation, and manually filtering the list of domains containing specific words, such as the names of Bulgarian cities.

The dataset **General News in English from Bulgaria** was automatically collected from various sources in Bulgaria: 140 web domains and 19,120 web pages. The total number of words in the collected General News in English is 18,631,384, distributed across 876,739 sentences.

The dataset **Filtered General News in English from Bulgaria** is drawn from the same 140 web domains and 19,120 web pages. However, the total number of words is lower: 5,512,392, distributed across 237,371 sentences, as the texts are further filtered to ensure that their content is specifically focused on Bulgaria.

The dataset **Filtered News in English about the Bulgarian EU Council Presidency from Bulgaria** was collected from 55 web domains and 1,402 web pages. The total number of words in the collected news in English is 504,596, distributed across 20,616 sentences.

The texts in the three datasets were aggregated into JSON files, which contain extracted metadata and automatic categorisation of the content into 185 domains. Automatic linguistic processing – sentence splitting and tokenisation – was performed.

The dataset **Collection of Bulgarian Texts** contains texts obtained mainly from the internet. Texts are organised into 66 files by style and thematic domain. A general classification into different styles (administrative, science, news, and fiction) is provided, and texts are further classified into thematic domains: Adventure, Archaeology, Architecture, Arts, Astronomy, Biology, Chemistry, Children, Computers, Court, Culture, Ecology, Economics, Education, Engineering, Entertainment, Geography, Health, History, Law, Lifestyle, Linguistics, Literature, Maths, Medicine, Military, Pedagogy, Philosophy, Physics, Politics, Psychology, Relations, Religion, Science Fiction, Science, Society, Sociology, Sport, Technology, Travel, and Unclassified. Automatic sentence splitting is applied; the text format is one sentence per line. The collection contains 28,919,379 sentences. The files are not supplemented with metadata and the sentence order is arbitrary.

The dataset **Collection of English Texts** contains texts obtained mainly from the internet. Texts are organised into 45 files by style and thematic do-

main. A general classification into different styles (administrative, science, news, and fiction) is provided, and texts are further classified into thematic domains: Court, Culture, Ecology, Economics, Health, History, Military, Physics, Politics, Science Fiction, Society, Technology, and Unclassified. Automatic sentence splitting is applied; the text format is one sentence per line. The collection contains 8,144,881 sentences. The files are not supplemented with metadata and the sentence order is arbitrary.

The project **Infrastructure for Fine-tuning Pretrained Large Language Models** (IfGPT)[12] provides an opportunity to collect, clean, and curate additional Bulgarian data. Table 5 shows the amount of new data collected. This includes older texts, both periodicals and books, which require further processing – OCR, pagination, and metadata extraction using the layout and structure of the texts (for example, metadata such as date and source in the header and footer of periodicals, publishing information from the title page, etc.)—as well as post-processing procedures for the extracted text, including the removal of boilerplate content and correction of hyphenation. The capabilities of LLMs have been tested for one or more of these tasks; in particular, OCR was combined with text completion (separating articles in newspapers) and metadata extraction using the Claude Sonnet 4.6 API.

Source	# texts	# tokens	Licence
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 5: New data added to IfGPT dataset.

The **Multilingual Image Corpus** (**MIC21**) is a recently developed dataset designed to advance research in multilingual and multimodal data processing (Koeva et al., 2022). It provides pixel-level annotations for over 203,000 objects in more than 21,000 images, covering 730 classes organised into four thematic domains and 130 subdomains.[13] These annotations support the development of specialised models for object detection, segmentation, and classification. The annotated object classes depicted in the images are structured within an Ontology of Visual Objects, enabling the construction of diverse datasets for a wide range of tasks. This ontological framework supports learning inter-object associations, identifying relationships between objects, and aligning objects and their relations with textual content. Class labels are enriched with synonyms, definitions, and usage examples in 25 languages, making the dataset suitable for applications such as multilingual image captioning, visual question answering, and multimodal machine translation. In a recent extension of the dataset, images have been accompanied by brief narrative de

[12] https://ifgpt.dcl.bas.bg/en/

[13] https://dcl.bas.bg/en/projects_list/mic21/

4. Extensive metadata description in IfGPT

Each document in the IfGPT dataset is described by a set of mandatory and optional metadata fields. The mandatory fields ensure consistent descriptions across all documents, covering text characteristics, domain information, and quantitative document statistics. The optional fields provide supplementary descriptive details where available, including authorship, stylistic properties, and task suitability. Most of the metadata originates from the BulNC metadata, which was structured as a graph (Koeva et al., 2016); however, there are some specific metadata fields related to LLMs, for example: **PersonallyIdentifiableInformation**, **BiasedInformation**, **LicenseLink**,**TaskCategories**.

4.1. Metadata types

The mandatory metadata includes: **Identifier** (unique document ID with language prefix `bg`), **Licence** (terms of use; various CC licence types, etc.), **PublicationDate** (original publication date), **DocumentTitle** (title of the document), **Source** (journal or website), **Medium** (modality: text, multimodal), **Url** (original web address), **Domain** (up to six thematic domains from a predefined list), **Keywords** (up to six descriptive terms), **NumberWords** (total word count), **NumberSentences** (total sentence count), **NumberParagraphs** (total paragraph count), **NumberTokens** (total token count), **PersonallyIdentifiableInformation** (an array of values for all sentences, calculated as the proportion of tokens flagged as PII), and **BiasedInformation** (an array of values for all sentences, calculated as the proportion of tokens flagged as potentially expressing bias).

The optional metadata includes: **Author** (name(s) of the text's creator(s)), **Style** (literary register, e.g. Legal, Journalism, Administrative), **Type** (document genre, e.g. book chapter, newspaper article, blog post), **Subdomain** (narrower thematic classification linked to a parent domain), **TranslatedDocument** (whether the document is original Bulgarian or a translation), **CollectionDate** (date of data collection in ISO 8601), **LicenseLink** (URL of the licence describing its terms), and **TaskCategories** (intended NLP applications from a predefined list, e.g. question answering).

An illustrative example of metadata organisation is shown on Figure 1.

4.2. Metadata management

To store and manage the metadata described above, a Neo4J graph database is used.[14] Graph databases are well suited to this purpose, as they are designed to handle large volumes of interconnected data efficiently, support horizontal scaling, and maintain performance even under complex queries (Francis et al., 2018). Neo4J is chosen for its high performance, native support for the Cypher query language, and extensive community ecosystem.

The metadata is organised according to a graph schema that captures the key entities and their interrelationships. Five node types are defined: **Document** nodes, which contain the core descriptive properties of each text (identifier, title, source, domain, author, licence, etc.); **Domain** nodes, characterised by a name and a parent category to support hierarchical thematic classification; **Author** nodes, storing author names and optional biographical details; **Source** nodes, recording the name and URL of the publishing organisation; and **Licence** nodes, defined by a single type property.

The relationships between these nodes are represented as directed edges: a **Document** belongs to a **Domain** (`BELONGS_TO`); a **Domain** may be a subcategory of another **Domain** (`SUBCATEGORY_OF`); a **Document** is licensed under a **Licence** (`LICENSED_WITH`); a **Document** is attributed to an **Author** (`WRITTEN_BY`); a **Document** is associated with a **Source** (`PUBLISHED_IN`); etc. Together, these nodes and edges form a flexible, queryable representation of the metadata that supports both document retrieval and downstream NLP tasks.

The metadata graph database currently contains a total of 237,795 documents, described through `metadata`, `Author`, `Domain`, and `Licence` nodes, connected by a number of relation types: `BELONGS_TO`, `LICENSED_WITH`, `WRITTEN_BY`, `PUBLISHED_IN`, etc.

The use of a graph database for managing metadata offers several key advantages. It models the rich relations interconnecting documents, domains, licences, sources, and more, making it much more expressive for metadata exploration. Complex queries combining multiple criteria are handled as efficient graph traversals (Figure 2 illustrates this process). Graph databases also scale efficiently when new relationship types are introduced (e.g. linking Keywords and Domain nodes) without requiring changes to the schema.

Domain distribution. The collection spans 45 thematic domains and subdomains. The most represented domains account for the majority of documents: Science (19.4%), Public Administration (15.9%), Economics (15.4%), and Politics (12.7%); 16.7% of texts have no assigned domain. The remaining 41 domains each account for less than 6% of documents, with several highly specialised domains, such as Architecture, Computer Science, and Physics.

Licensing. The vast majority of documents are openly licensed (93.4%), with the most common licences being CC-BY-SA (42.8%) and CC0 (45.9%). Only 6.6% of documents have a restricted licence.

Authorship. Author metadata is available for 93.7% of documents, while 6.3% lack authorship information. The collection references 4,563 distinct authors across all documents, including both individuals and organisations.

Document statistics. The collection currently available for searching comprises approximately 718.4 million words and 866 million tokens. On average, each document contains 3,642 tokens (3,021 words) in 207 sentences, indicating that the collection predominantly consists of full-length texts.

Time period coverage. Publication dates range from 1935 to 2022. The bulk of the collection is concentrated in the period 2008–2011, which accounts for approximately 55% of all documents. Current efforts aim to provide more data covering the pre-2000 (pre-digital) period.

5. Ensuring linguistic integrity of data

Ensuring the linguistic integrity of the data is a major priority in compiling the IfGPT dataset, as poor content quality has been shown to directly affect model performance and introduce systematic biases (Kreutzer et al., 2022). To address this, we have implemented two main procedures: deduplication and the removal of unsuitable and malformed texts.

Deduplication is carried out in two steps. First, metadata attributes such as source, publication year, domain, title, and author are used to identify and remove identical texts appearing in multiple text collections, which substantially reduces the computational burden of subsequent processing. The core deduplication procedure is then performed using MinHash combined with Locality Sensitive Hashing (LSH) (Lee et al., 2022), which detects both exact and near-duplicate texts by estimating N-gram overlap across document pairs.

The second procedure involves identifying and removing boilerplate (e.g., navigation menus, repeated footer text, legal disclaimers), correcting typographical errors, removing incomplete or malformed sentences, and filtering harmful, offensive, and toxic content.

[14] https://neo4j.com/

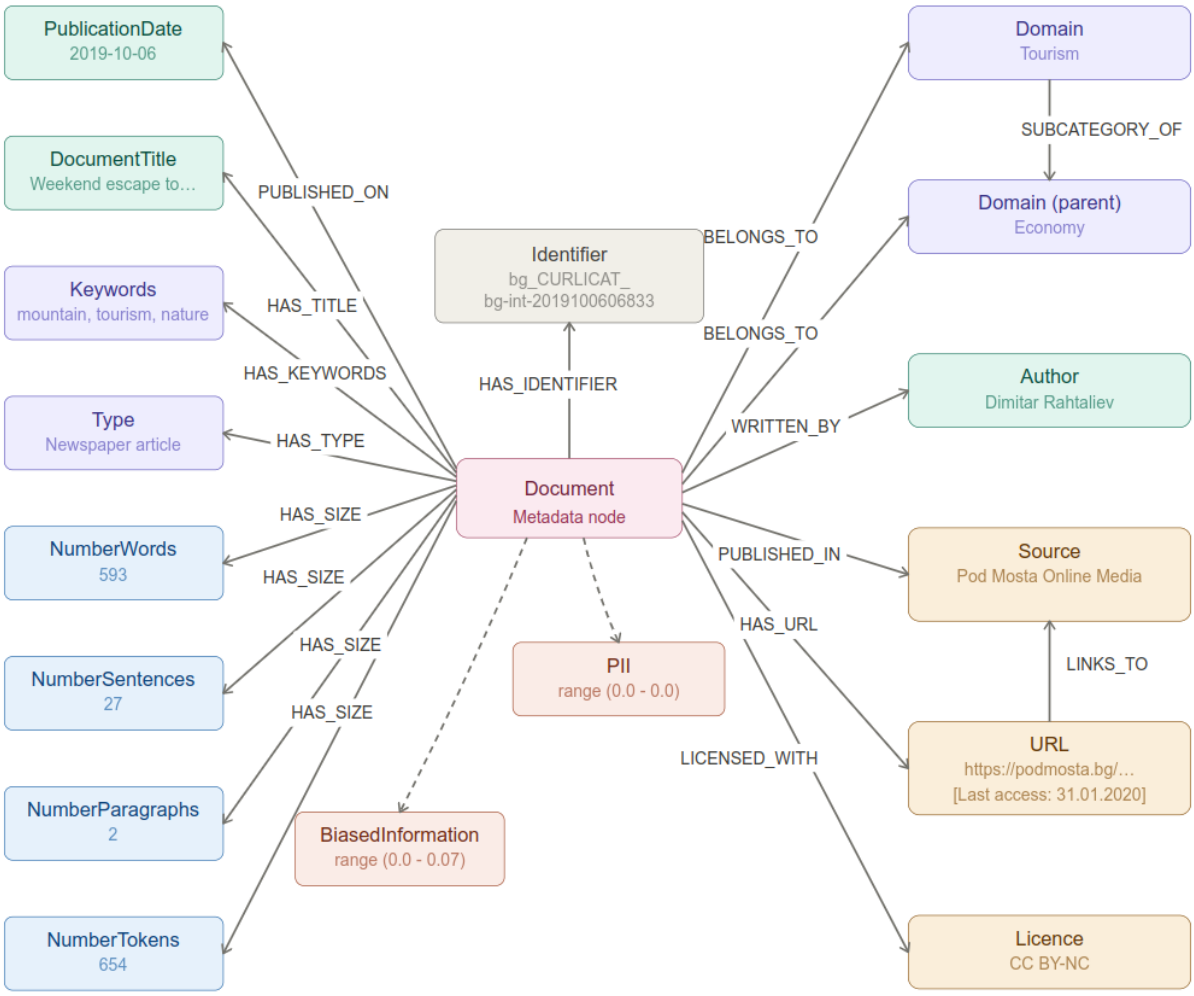

Figure 1: IfGPT Dataset Metadata.

The PII and BiasedInformation metadata fields are also directly relevant to the quality assessment and maintenance of the data. Instead of being discarded, potentially sensitive information and biased content are flagged, allowing users to apply their own filtering criteria depending on the task at hand.

6. Access

The IfGPT dataset is publicly accessible through a dedicated web interface,[15] which allows users to search and browse the complete collection of documents. The interface offers four filtering mechanisms (see Figure 3).

Licence filter. Users can restrict results by licence type, selecting either general or specific licences, e.g. all Creative Commons or specific licences, other open licences, and can include or exclude data with restricted licences.

Domain filter. Documents can be filtered by one or more domains.

Period filter. Users can specify a publication date range by setting a start and/or end year to restrict results to a particular period (omitting either sets it to the default, i.e. the earliest or latest document year in the database).

Keywords filter. Free-text keyword search is supported, with multiple keywords accepted as a comma-separated list.

Search results are displayed as paginated document cards, each showing the document title, domain tags, licence, document type, publication date, source, URL to the original source, and quantitative properties including the number of paragraphs, sentences, and words (see Figure 3).

The interface also provides three download options for the retrieved results: (1) metadata of retrieved documents (JSON format), (2) list of links to original sources (TSV format), and (3) full data (link to a ZIP archive) subject to confirmation of details and agreement to the Terms and Conditions for download, including the restrictions imposed by the licences of the original documents.

[15] Available at https://ifgpt.dcl.bas.bg/ifgpt-dataset/

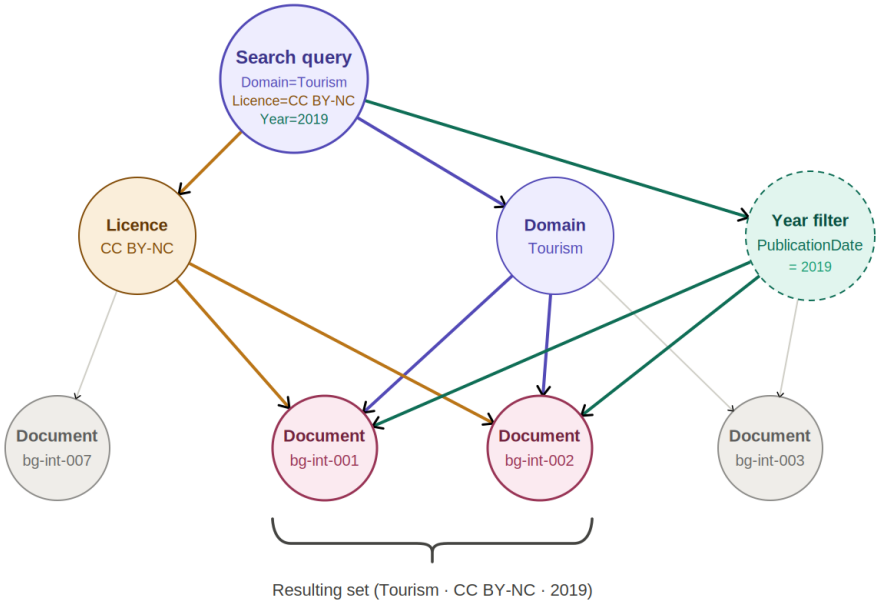

Figure 2: IfGPT Dataset Search in the graph database.

7. Conclusion

By describing the IfGPT dataset, we demonstrate that the reference corpora are suitable for inclusion in LLM datasets, as they share several important similarities in design, compilation, and management. Both BulNC and IfGPT aim to represent authentic Bulgarian language and include multiple genres, domains, and text types, as both linguistic research and language model training require exposure to varied language. Both resources also require cleaning and deduplication, filtering out incomplete or malformed texts. Both are stored in standard formats, divided into documents, paragraphs, and sentences. Although the IfGPT dataset does not provide as extensive linguistic annotation as BulNC, such annotation could also be implemented. This means that both resources can complement each other: the BulNC is part of the IfGPT dataset, but appropriate parts of the IfGPT dataset that fulfil balance and distribution requirements can also be added to the Bulgarian National Corpus. Intensive work on developing tools for detecting bias and PII in IfGPT texts will also be useful for application to BulNC documents.

Unlike some LLM datasets, the IfGPT dataset, inheriting from BulNC, possesses extensive metadata descriptions. The BulNC metadata, originally organised as a graph, is enhanced with some LLM-specific metadata such as PII and bias scores, licence information, and is further managed as a graph database. The common metadata scheme for both resources is beneficial in two ways: searching through the metadata for relevant texts from IfGPT for BulNC, and adding new relevant texts simultaneously to BulNC and IfGPT.

Future developments include: (a) adding new and diverse text data to both resources; (b) expanding metadata descriptions, especially for texts for which some metadata categories are not assigned values; (c) validating and improving text data quality in both resources; and (d) enhancing accessibility and providing easy access to the IfGPT data for various purposes.

8. 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 4.3.1) [Windows, macOS, Linux]. Corpus analysis toolkit.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv*.

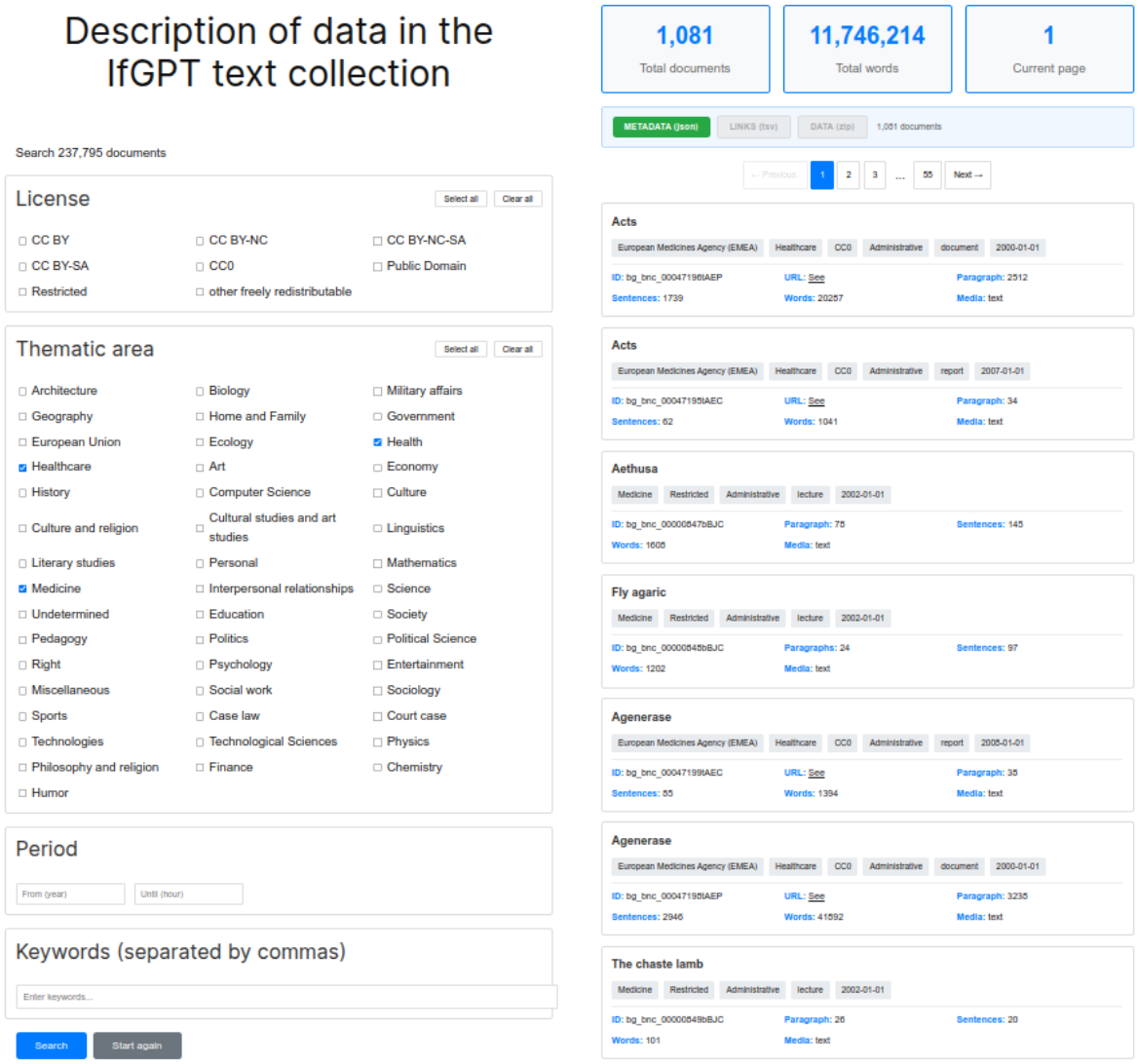

Figure 3: Web interface for searching and selecting datasets from IfGPT. Results from the search are shown on the right

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

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, Portorož, Slovenia. ELRA.

Jesse Egbert, Douglas Biber, and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge University Press, Cambridge.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darundefinedis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. ParlaMint II: advancing comparable parliamentary corpora

across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*.

Brett J. Hashimoto and Mike Nelson. 2024. Recent trends in corpus design and reporting: A methodological synthesis. *Research in Corpus Linguistics*, 12(1):59–88.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. Lexical Computing Ltd. / Masaryk University, Brno.

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.

Jan Kocoń, Mateusz Kopeć, et al. 2025. CLARIN-PL: A user-centred language technology infrastructure. *Language Resources and Evaluation*.

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, 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.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France. ELRA.

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.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia. ELRA.

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.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of LLM agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Leni Fowl, et al. 2023. ROOTS: A multilingual annotated pretraining corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 369–380, Toronto, Canada. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and pragmatic multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.

Piotr Pęzik, Paweł Kowalczyk, Łukasz Dróżdź, and Paweł Wilk. 2016. PELCRA for National Corpus of Polish: Search Engine 2. CLARIN-PL Digital Repository.

Pavel Rychlý. 2007. Manatee/Bonito: A Modular Corpus Manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pages 65–70.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Ian Magnusson, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. Stanford.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey. ELRA.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiş, Dan Tufiş, Radovan Garabík, Simon Krek, Andraž Repar, Matjaž Rihtar, and Janez Brank. 2020. The MARCELL legislative corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. ELRA.

Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartłomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Dan Tufiş, Radovan Garabík, Simon Krek, and Andraž Repar. 2022a. Introducing the CURLICAT corpora: Seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 100–108, Marseille, France. ELRA.

Tamás Váradi, Marko Tadić, Svetla Koeva, Maciej Ogrodniczuk, Dan Tufiş, Radovan Garabík, Simon Krek, and Andraž Repar. 2022b. Curated multilingual language resources for CEF AT (CURLICAT): Overall view. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 341–342, Ghent, Belgium. European Association for Machine Translation.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.