A New Massive Multilingual Dataset for High-Performance Language Technologies

Ona de Gibert[1], Graeme Nail[2], Nikolay Arefyev[3], Marta Bañón[4],
Jelmer van der Linde[2], Shaoxiong Ji[1], Jaume Zaragoza-Bernabeu[4]
Mikko Aulamo[1], Gema Ramírez-Sánchez[4], Andrey Kutuzov[3],
Sampo Pyysalo[5], Stephan Oepen[3] and Jörg Tiedemann[1]
University of Helsinki, Finland[1] University of Edinburgh, UK[2] University of Oslo, Norway[3]
Prompsit, Spain[4] University of Turku, Finland[5]
{ona.degibert, shaoxiong.ji, mikko.aulamo, joerg.tiedemann}@helsinki.fi[1],
{graeme.nail, jelmer.vanderlinde}@ed.ac.uk[2],
{nikolare, andreku, oe}@ifi.uio.no[3]
{mbanon, jzaragoza, gramirez}@prompsit.com[4],
sampo.pyysalo@utu.fi[5]

Abstract

We present the HPLT (High Performance Language Technologies) language resources, a new massive multilingual dataset including both monolingual and bilingual corpora extracted from CommonCrawl and previously unused web crawls from the Internet Archive. We describe our methods for data acquisition, management and processing of large corpora, which rely on open-source software tools and high-performance computing. Our monolingual collection focuses on low- to medium-resourced languages and covers 75 languages and a total of ≈ 5.6 trillion word tokens de-duplicated on the document level. Our English-centric parallel corpus is derived from its monolingual counterpart and covers 18 language pairs and more than 96 million aligned sentence pairs with roughly 1.4 billion English tokens. The HPLT language resources are one of the largest open text corpora ever released, providing a great resource for language modeling and machine translation training. We publicly release the corpora, the software, and the tools used in this work.

Keywords: Parallel Corpus, Monolingual Corpus, Low-resource Languages, Pre-training Datasets

1. Introduction

The development of Large Language Models (LLMs) pre-trained on ever-increasing amounts of text combined with the ongoing advancements in Machine Translation (MT) has made the need for vast amounts of high-quality textual data more pressing than ever. Since the acquisition of large text corpora is a challenge, most works focus on the pre-processing of previously released corpora with new methods, such as more strict textual filters or removal of biased or explicit content. In this work, we present a massive, brand-new dataset for language modeling and MT training based on web crawls produced by the Internet Archive,[1] used for the first time at this scale to create multilingual text corpora, and from CommonCrawl.[2]

Under the umbrella of the High Performance Language Technologies (HPLT) project[3] (Aulamo et al., 2023), we obtained access to the web crawls (1.85 PB of data in total at the current stage of the project), downloaded and processed them to create monolingual and parallel corpora with rich metadata: the HPLT language resources. We release the collection under the permissive CC0 license[4] through our project website[5] and OPUS[6] (Tiedemann, 2012). We also publish open-source tools and pipelines used for processing huge web archive data packages so that our real use case can serve as an example for others inside and outside the research community. Software and tools are released through GitHub.[7]

Our contributions can be summarized as follows:

- *monoHPLT*: Monolingual datasets covering 75 languages and over 5.6 trillion tokens.
- *biHPLT*: Parallel datasets covering 18 language pairs and over 96 million sentence pairs.
- *multiHPLT*: Synthetic datasets obtained by pivoting our parallel datasets through English covering 171 language pairs and 157 million sentence pairs.
- *Bitextor* (Esplà-Gomis et al., 2016) models: 22 MT models for fast translation and bilingual document alignment covering 9 languages.

[1] https://archive.org/
[2] https://commoncrawl.org/
[3] https://hplt-project.org/

[4] We do not own any of the text from which these text data has been extracted. We release the data under a specific takedown policy, where any user can ask us to remove their data.
[5] https://hplt-project.org/datasets/
[6] https://opus.nlpl.eu/
[7] https://github.com/hplt-project

LREC-COLING 2024, pages 1116–1128

- *Bicleaner AI* (Zaragoza-Bernabeu et al., 2022) models: 9 new Bicleaner models for sentence pair scoring.
- Scripts and tools for managing, downloading and processing large amounts of web-crawled corpora.

The rest of the paper is organized as follows. Section 2 provides an overview of previous work in constructing corpora for pre-training. Section 3 describes the acquisition of the presented resources. Section 4 presents in detail the introduced language resources. Finally, Section 5 concludes our work and discusses future lines of research.

2. Related Work

The development of LLMs and highly multilingual MT systems demands large amounts of high-quality data. The scale of training data required by these models makes it effectively impossible to only use curated samples; instead, the common solution to gathering sufficient data is to source it from the Internet. The compilation of text corpora from the Web, both monolingual and bilingual, has been going on for a long time (Kilgarriff and Grefenstette, 2003). While some noteworthy efforts focus on language-specific curated datasets, such as C4 in English (Dodge et al., 2021) and WuDaoCorpora in Chinese (Yuan et al., 2021), the current capacity of models in the field has grown, leading to a move towards large multilingual collections.

Regarding monolingual resources, one of the most used sources is CommonCrawl (CC), produced by a non-profit organization that has published a collection of monthly multilingual web snapshots since 2011. Due to its size and noisy nature, there have been multiple efforts at processing CC data to compile cleaned versions: the multilingual OSCAR corpus (Suárez et al., 2019), as well as the English corpora Pile-CC (Gao et al., 2020), C4 (Dodge et al., 2021) and its multilingual counterpart mC4 (Xue et al., 2021). Other well-known multilingual corpora for language modeling include the recent BigScience ROOTS Corpus (Laurençon et al., 2022), covering 59 languages from a diverse set of sources, CuturaX (Nguyen et al., 2023), a cleaned multilingual dataset in 167 languages, MADLAD-400 (Kudugunta et al., 2023), a large audited dataset in 419 languages, Glot500 (Imani-Googhari et al., 2023), a corpus covering 511 languages, and SERENGETI (Adebara et al., 2023), a dataset in 517 African languages. Bapna et al. (2022) built a massively multilingual dataset in over 1,500 languages; however, they did not release it publicly.

For parallel corpora, the largest publicly available bitext collection is OPUS (Tiedemann, 2012). The collection includes several large multilingual corpora, such as Paracrawl (Bañón et al., 2020), its current version 9 covers 42 languages and English-centric sentence pairs; CCMATRIX (Schwenk et al., 2021), obtained from CC, and the recent NLLB data (Costa-jussà et al., 2022), which aims at covering as many language pairs as possible.

When dealing with web-crawled corpora, concerns arise regarding the original sources of the data and its level of noisiness. Several works have addressed this issue (Kreutzer et al., 2022; Abadji et al., 2022) and have led researchers to further explore their own datasets and develop new metadata schemes, such as adding genre labels (Laippala et al., 2022; Kuzman et al., 2023), or to include extended annotations such as length, noise and adult content tags (Abadji et al., 2022). The HPLT language resources also contain additional paragraph-level metadata; see subsection 4.1 for more detail.

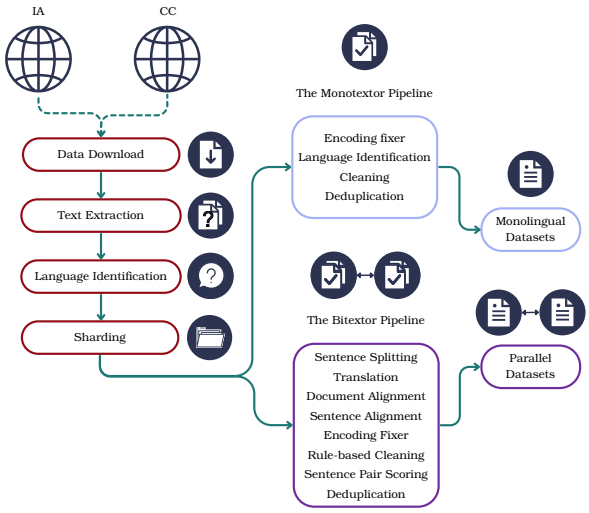


Figure 1: General overview of the HPLT acquisition and processing pipeline.

3. From Raw Data to Refined Corpora

The management and processing of large datasets both introduce their separate challenges. In this section, we provide a detailed account of the methods, techniques, and considerations employed to collect the raw data and transform it into the corpora presented in Section 4. A general overview of the pipeline is depicted in Figure 1.

Data download Data acquisition in HPLT relies on two main sources of web crawls: the Internet Archive and Common Crawl. The national High-Performance Computing (HPC) storage resources of Sigma2[8] and CESNET[9] were used to down-

[8] https://www.sigma2.no/data-storage
[9] https://www.cesnet.cz/

Crawl (collection)	CC40	IA WIDE15	IA WIDE16	IA WIDE17	Total
# WARC files	80 000	361 431	754 143	662 381	1 857 955
# files after `warc2text`	384 360	1 490 152	1 955 584	2 403 058	6 233 154
Compressed text size, TB	8.4	19	42	18	87.4
Uncompressed text size, TB	18.04	38.15	130.82	43.65	230.7
# text files	127 853	495 512	977 792	798 811	2 399 968

Table 1: Sizes of the raw texts extracted from crawls. 'CC' stands for 'Common Crawl', 'IA' stands for 'Internet Archive'.

load and pre-process web crawls from these two sources. The downloading scripts are published in the HPLT git repository.[10] These enable parallelized data downloading while automatically verifying and retrying failed downloads after a back-off period. These features are vital for downloading large file collections such as web crawls.

For the current data release, we have downloaded three large web crawls from the Internet Archive (IA) named WIDE15, WIDE16 and WIDE17, along with the CC-MAIN-2022-40 (CC40) crawl from Common Crawl. These crawls occupy a total of 1850 TB and are stored in WARC (Web Archive) format[11]. More data will be made available in the future releases.

Text Extraction WARC files contain many types of data besides written text: images, sound, video, etc. In order to extract raw texts and conduct preliminary language identification, the downloaded crawls were processed by the `warc2text` tool from the Bitextor pipeline.[12] `warc2text` finds documents containing text in some natural language and does fast preliminary filtering of undesirable documents based on their URL or HTML tags. More thorough filtering happens at the next stages. From the remaining documents, it extracts raw unformatted text and performs initial, document-level language detection. Running whitespace is normalized, and paragraph-like segments, as defined by HTML block elements (`<p>`, `<ul>`, `<ol>`, etc.) are encoded as newlines in this raw text. The output of `warc2text` consists of compressed base64-encoded raw texts along with the URLs of the original web pages these texts originate from. This data is grouped into directories by language, which is detected using the CLD2 language classifier[13]. Table 1 presents summary statistics for the crawls, showing that out of the four sources, WIDE16 produces by far the largest amount of text. In this step, we obtained 87.4 TB of compressed or 230.7 TB of uncompressed text in total.

After text extraction, we selected 77 languages with the highest amount of obtained raw text for this data release. We plan to add more languages in the following releases. The volume of uncompressed text obtained differs significantly across languages, from 2.2 GB for text classified by CLD2 as Esperanto to 77.5 TB for English, while the number of documents has a minimum of 314K for Pashto and a maximum of 12.8B for English. For most languages, the majority of texts come from the largest crawl, WIDE16; however, for Chinese, the main source is WIDE17, while Esperanto, Basque, and Nepali primarily come from CC40, despite it containing only a fifth of the text that WIDE16 does. Thus, a combination of different crawls, including small ones, seems to be beneficial for good coverage of different languages. The source crawl distribution per language can be consulted in Appendix A.

Language Identification The preliminary per-document language identification employs the CLD2 language identifier described above as the fastest solution. It is conducted as a part of `warc2text` processing. However, at a later stage of data processing (see below), we use FastSpell[14] (Bañón et al., 2024) for more accurate language identification at the level of paragraph-like segments.

Sharding To better deal with the amount of data to be processed, we organise the raw text records into 256 shards. The Bitextor pipeline identifies parallel text within a single shard, [15] and therefore, records are placed into shards by their domain name, excluding the top-level domain, to increase the likelihood of matches. Since the distribution of data is not uniform across shards, we batch the data into equally sized chunks for each shard to help balance the computational requirements.

[10] https://github.com/hplt-project/ia-download

[11] https://www.iso.org/standard/68004.html

[12] https://github.com/bitextor/warc2text

[13] https://github.com/CLD2Owners/cld2

[14] https://github.com/mbanon/fastspell

[15] https://github.com/bitextor/bitextor/blob/master/docs/CONFIG.md#preprocessing-and-sharding

For monolingual text extraction, the division among shards is ignored.

After these steps are complete, the monolingual and bilingual text processing pipelines go on separately, as described below. The following steps have been performed on the EuroHPC cluster LUMI[16].

3.1. Monolingual Text Processing

After the sharding step, we process the monolingual extracted text with cleaning tools in order to perform fixes at character level and to enrich the corpora with additional metadata that can be used to produce filtered versions for different applications.

The Monotextor Pipeline To be able to process 124TB of compressed text and scale across many compute nodes in an HPC cluster, a new pipeline based on Slurm scripts using the existing Monotextor tool was developed.[17] The pipeline performs the following steps:

1. **TSV Formatting**: for each shard, a tab-separated file is created where each line contains a document URL, a text paragraph, and a collection name. The file is split into batches of equal amounts of uncompressed text to balance subsequent processing jobs. For the following steps, each batch file is processed in parallel across the number of lines with GNU Parallel (Tange, 2023).
2. **Monofixer**[18]: every line, containing a paragraph-like segment of text, is processed by the character and encoding fixer, including fixing mojibake (encoding errors), unescaping HTML entities and removing HTML tags.
3. **FastSpell**: We perform language identification at paragraph-level in two steps. First, each paragraph receives a language tag given by fastText. Then, we refine the language identification using Hunspell dictionaries for improved precision. We check the paragraph for spelling errors with Hunspell dictionaries based on a list of similar languages[19] to the one identified by fastText. The language whose dictionary produces the least spelling errors is the final prediction.
4. **Monocleaner**[20]: each paragraph is also assigned a fluency score, computed with a 7-gram modified Knesser-Ney character language model. Each language model (one per language) is trained on samples of about 200,000 sentences mostly coming from the monolingual part of OPUS corpora. Only corpora coming from non-web-crawled data and languages that have not been automatically identified are chosen. Data from Wikipedia dumps are used for languages that do not have enough data in OPUS. This fluency score can be used to estimate the 'quality' of paragraphs in the document, allowing to filter out noise that may be detrimental for training language models.
5. **JSON formatting**: Finally, each batch tab-separated file is converted to JSON-lines (JSONL) format.

Language Mappings In addition to the processing described above, some modifications to how languages are stored after WARC text extraction have been made:

- CLD2 uses old Hebrew '`iw`' language code, so it has been renamed to use the official '`he`'.[21]
- Norwegian Bokmål is identified as '`no`' by CLD2, so it has been changed to '`nb`' to avoid possible confusion, as '`no`' may also refer to all the Norwegian variants, not only Bokmål.
- For consistency with the rest of the languages, where we are not separating by writing script, traditional and simplified Chinese ('`zh-Hant`' and '`zh-Hans`') have been merged into '`zh`'.
- For the monolingual collection, Serbo-Croatian languages (Bosnian '`bs`', Croatian '`hr`' and Serbian '`sr`') have been merged under the '`hbs`' code. Because of their mutual intelligibility, these languages are often mixed up with each other during language identification.

This process leaves **75 languages** for the monolingual data release.

De-duplication De-duplication of training corpora is of utmost importance, especially in the case of web-crawled text collections, which can often contain multiple copies of the same text appearing on different web pages. At the same time, overly aggressive de-duplication can lead to biased corpora, which are not representative of frequency patterns in the corresponding languages anymore. The datasets we release aim to allow end users to

[16] https://www.lumi-supercomputer.eu/

[17] https://github.com/hplt-project/monotextor-slurm

[18] https://github.com/bitextor/bifixer

[19] https://github.com/mbanon/fastspell/blob/main/src/fastspell/config/similar.yaml

[20] https://github.com/bitextor/monocleaner

[21] https://www.loc.gov/standards/iso639-2/php/langcodes_name.php?iso_639_1=he

decide whether they would like to apply any additional pre-processing.

For these reasons, we limited ourselves to removing near-duplicates **on the document level**, using a variation of the MinHash algorithm (Broder, 2000). This removed approximately 70-80% of the original data for high-resource languages and 40% for low-resource languages. In total, after de-duplication, the monolingual dataset was reduced to nearly a third of the size of the original (from 21 TB to 7.5 TB), but at the same time much more balanced and suitable for training language models. Note that we also release the data *before de-duplication* to preserve the possibility to reproduce or refine our de-duplication pipeline.

The final statistics and data format of the monoHPLT corpora are described in subsection 4.1 below.

3.2. Bitext Extraction

The sharded monolingual data is the input to the bitext extraction pipeline used to create parallel corpora. We rely on previous experience from the ParaCrawl[22] and MaCoCu projects[23] and adjust tools and procedures from the Bitextor pipeline[24] according to the needs and languages in the HPLT setup.

In this release, we focus on English-centric data as we expect the largest potential outcome of parallel data from the alignment to English. Furthermore, the Bitextor pipeline relies on automatic document translation in one of the steps and the performance of translations into English is more reliable than translations into other languages, especially for lesser-resourced languages. The initial release covers 18 language pairs with a strong focus on lesser-resourced languages and a few non-European languages to increase the diversity of parallel data available for MT development.

The Bitextor Pipeline The bitext extraction pipeline is based on Bitextor.[25] We extended scripts developed for ParaCrawl (Bañón et al., 2020) for scheduling and workflow automation to meet our needs. We adjusted and further developed the pipeline for the needs of HPLT on the LUMI supercomputer. Mining bilingual sentence pairs using this pipeline and English as one of the languages consists of the following processing steps for each language pair:

1. **Sentence Splitting**: splits the documents into sentences using a language-specific sentence splitter. When there is no language-specific sentence splitter, default to English.

2. **Translation**: translate the non-English sentences into English for document alignment. For these steps, we needed to develop fast MT models described below. MarianNMT (Junczys-Dowmunt et al., 2018) was used for automatic translation and adapted to work with the AMD GPUs available on LUMI.[26]

3. **Document Matching**: match English and translated documents using TF/IDF. Code was improved[27] to work with less memory to better handle larger batch sizes.

4. **Sentence Alignment**: match English and translated sentences in the matched documents using Bleualign[28] (Sennrich and Volk, 2010), which relies on the English translated sentence and the original English sentence.

5. **Bifixer** (Ramírez-Sánchez et al., 2020): fix encoding and orthographic issues, similar to Monofixer for monolingual text data.

6. **Bicleaner-hardrules** (Ramírez-Sánchez et al., 2020): remove noisy sentence pairs for obvious noise based on rules, poor language identified using FastSpell and vulgar language based on specific language modelling.

7. **Bicleaner AI** (Zaragoza-Bernabeu et al., 2022): score sentence pairs to indicate whether they are mutual translations (with a value near to 1) or not (with a value near to 0). We keep sentences whose Bicleaner score is above 0.5.

8. **Reduce**: collect and concatenate all data across shards and collections. In our case, this is a combination of all the data extracted from CC-40, WIDE15, WIDE16 and WIDE17.

9. **De-duplication and TMX Formatting**: the final step generates a TMX file. In this step, the sentence pairs are de-duplicated, ignoring differences in punctuation. The source URLs are retained so that a single sentence pair can have multiple URLs, identifying all the documents that it occurred in.

[22] https://www.paracrawl.eu/
[23] https://macocu.eu/
[24] https://github.com/paracrawl/cirrus-scripts/tree/lumi
[25] https://github.com/bitextor/bitextor
[26] https://github.com/hplt-project/lumi-marian
[27] https://github.com/hplt-project/document-aligner
[28] https://github.com/bitextor/bleualign-cpp

3.2.1. Bitextor Models

Document matching in the Bitextor pipeline requires the translation of one language into the other in order to use efficient monolingual matching strategies to find parallel documents in the vast space of extracted texts. This requires efficient translation models to make it computationally feasible to process data at the scale involved in this work. Bitextor already supports a number of languages from prior work, but their coverage is limited. OPUS-MT[29] provides additional resources in terms of pre-trained models that can be employed directly for translation or for distillation, as detailed below.

For this data release, we trained new efficient student MT models to enable the extraction of additional language pairs. We adopted larger transformer-based MT systems as teacher models and distilled knowledge from the teacher to train student models and improve efficiency via sequence-level knowledge distillation (Kim and Rush, 2016). This technique allows the student model to learn from the teacher model to create a model of comparable quality but improved throughput thanks to its smaller size. We trained two different-size student models, `base` and `tiny`, to cover different quality-throughput requirements. We trained models for languages including `ar`, `ca`, `eu`, `gl`, `hi`, `jp`, `sw`, `vi`, and `zh` (in both simplified and traditional scripts, as well as a joint model). We release the student models under our GitHub repository[30] .

3.2.2. Bicleaner AI Models

For bilingual data filtering, we use Bicleaner AI, which aims to detect noisy sentence pairs in a parallel corpus. It indicates the likelihood of a pair of sentences being mutual translations (with a value near to 1) or not (with a value near to 0). Sentence pairs considered very noisy are scored 0.

Although there are already Bicleaner models available, we trained new Bicleaner models for the language pairs that we include in this release: `ar`, `eu`, `gl`, `he`, `hi`, `jp`, `sw`, `vi`, and `zh`. We have increased the total amount of language pairs available from 36 to 45,[31] also including many changes and improvements to the tool since version 1.0.1 made for ParaCrawl[32]. We make all of the newly introduced models available for download.[33]

[29] https://github.com/Helsinki-NLP/OPUS-MT

[30] https://github.com/hplt-project/bitextor-mt-models

[31] https://huggingface.co/models?other=bicleaner-ai

[32] https://github.com/bitextor/bicleaner-ai/blob/v2.3.2/CHANGELOG.md

[33] https://github.com/bitextor/bicleaner-ai/#download-a-model

4. The HPLT Language Resources

We next present the HPLT language resources, a new massive multilingual dataset covering 75 monolingual and 18 bitext corpora. We release all our collections under the permissive `CC0` license through our project website and OPUS.

4.1. Monolingual Datasets

Our monolingual collection covers 75 languages. We include high-resource languages such as English (en), Chinese (zh), Russian (ru) and Japanese (jp), as well as low-resource ones such as Esperanto (eo), Pashto (ps), Tatar (tt) and Welsh (cy). The full statistics for all languages are presented in Appendix B.

In total, after de-duplication, we release a collection of 5.25 billion documents (approximately corresponding to web pages), totaling 50.1 TB of uncompressed texts and approximately 5.6 trillion whitespace-separated word tokens. Figure 2 shows the proportions of language families and the largest individual languages in the released data.

We again emphasize that these web-derived corpora have only undergone essential pre-processing (see above), but no boilerplate removal, fine-grained filtering or extensive cleaning. At the same time, the texts are provided with metadata, which can be employed by end users to conduct their own filtering.

The datasets come as compressed `JSONlines` files, where each line is a valid JSON value representing a full document with metadata:

```
{"id":1, "document_lang":"en",
"scores":["0.76","0.70"],
"langs":["en","en"],
"text":"this is paragraph1\nparagraph2",
"url":"url1", "collection":"collection1"
}
```

The document text is in the 'text' field; paragraph-like segments are concatenated using newline separators (here, we have two paragraphs). The 'langs' and 'scores' fields contain lists with one entry per paragraph. The first corresponds to the paragraph languages identified by FastSpell (both paragraphs are in English here), and the second corresponds to the Monocleaner fluency score of each paragraph (in this case, the first paragraph is slightly closer to 'regular' English than the second one). The 'url' field provides the original URL from where the document was downloaded, and the 'collection' field features the identifier of a specific web crawl where the document was found (for example, 'WIDE16').

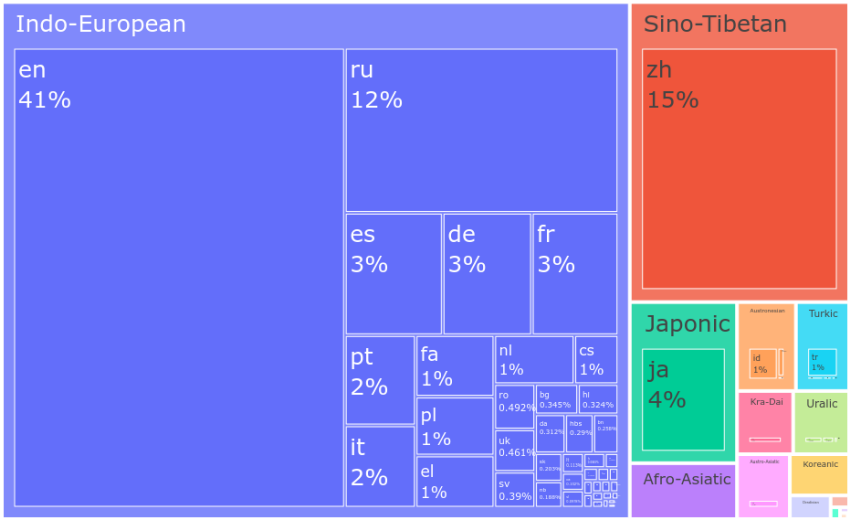


Figure 2: Size distribution for the monolingual corpora, organized by language family and language. The volume of texts ranges from 1.0 GB for text classified by CLD2 as Esperanto to 20.3 TB for English, accounting for 41% of the whole collection.

	Raw		Filtered		De-duplicated	
Language Pair	**# Segments**	**# Tokens**	**# Segments**	**# Tokens**	**# Segments**	**# Tokens**
Norwegian (nn)	28 701 601	496 496 331	649 435	6 308 500	132 538	2 082 878
Bosnian* (bs)	26 998 901	521 626 621	1 426 670	12 439 348	240 012	2 705 525
Basque (eu)	20 830 243	400 262 771	3 087 453	31 739 210	610 687	9 964 617
Maltese (mt)	135 103 434	2 820 798 439	9 170 421	133 140 189	854 820	18 819 145
Gaelic (ga)	101 001 090	2 013 971 167	15 644 170	144 323 574	994 746	16 327 484
Galician (gl)	56 101 411	1 015 559 754	5 789 361	49 604 655	1 063 103	13 904 758
Macedonian (mk)	91 293 129	1 868 196 128	20 474 476	221 370 998	1 139 051	18 562 461
Albanian (sq)	253 098 546	5 819 014 143	16 729 596	144 732 656	1 655 958	25 831 054
Swahili (sw)	247 557 313	5 746 490 123	24 448 577	209 062 077	1 710 205	20 039 612
Icelandic (is)	170 419 019	3 266 074 902	28 149 571	262 486 823	2 148 854	29 493 241
Serbian* (sr)	754 277 462	14 249 438 714	60 482 286	586 909 655	4 643 025	67 063 293
Chinese (zh)	530 119 983	9 162 123 041	47 852 076	510 404 638	5 306 570	83 811 653
Estonian (et)	865 431 226	15 476 948 993	72 976 009	752 767 471	6 089 791	95 943 562
Catalan (ca)	402 492 626	8 034 120 323	88 434 510	882 436 335	8 905 889	141 859 163
Croatian* (hr)	895 785 142	16 565 285 999	128 145 132	1 165 895 906	9 310 275	138 360 666
Hindi (hi)	1 043 856 525	19 246 270 565	117 341 153	996 036 740	12 043 069	165 139 713
Arabic (ar)	1 545 148 805	33 199 212 426	277 864 501	2 307 727 128	14 645 128	239 377 462
Finnish (fi)	3 826 974 191	65 312 092 463	495 310 671	4 186 819 006	25 176 462	338 063 309
Total	10 995 190 647	205 213 982 903	1 413 976 068	12 604 204 909	96 670 183	1 427 349 596

Table 2: Statistics on the extracted bitexts without filtering (Raw), after cleaning (Filtered) and after de-duplication (De-duplicated) ordered by available clean de-duplicated segments. All statistics are measured from the English side of each language pair. The symbol * indicates that a joint Bicleaner AI model has been used for processing those languages. The dashed line marks the boundary of 1 million clean and de-duplicated segments, which is often used as a threshold to distinguish low-resource and higher-resource languages.

4.1.1. Cleaned Version

In addition to the de-duplicated version of the monolingual datasets (v1.1), we also published the so called 'cleaned' version (v1.2).[34] In it, we removed full documents which satisfied at least one of the following 5 criteria:

1. URL is in the UT1 blocklist of adult sites.[35]
2. less than 5 words per segment (line) on average.

[34] https://hplt-project.org/datasets/v1.2

[35] https://dsi.ut-capitole.fr/blacklists/

```
<tu tuid="443" datatype="Text">
    <prop type="collection">wide15</prop>
    <prop type="collection">wide17</prop>
    <prop type="score-bicleaner">1.000</prop>
    <prop type="type">1:1</prop>
    <tuv xml:lang="en">
     <prop type="source-document">https://en.wikipedia.org/wiki
         /Criticism_of_Esperanto</prop>
     <prop type="source-document">https://en.wikipedia.org/wiki
         /Esperanto_language</prop>
     <prop type="source-document">https://en.m.wikipedia.org/wiki
         /Esperanto</prop>
     <prop type="checksum-seg">8d7b2e6e2780b0bb</prop>
     <seg>The number of speakers grew rapidly over the next few decades,
         at first primarily in the Russian Empire and Central Europe, then
         in other parts of Europe, the Americas, China, and Japan.</seg>
    </tuv>
    <tuv xml:lang="ca">
     <prop type="source-document">https://ca.wikipedia.org/wiki/Esperanto
         </prop>
     <prop type="checksum-seg">397455522023d177</prop>
     <seg>El nombre de parlants va créixer ràpidament en les primeres
         dècades, sobretot a l'Imperi Rus i a Europa Oriental; després a
         Europa Occidental, a les Amèriques, a la Xina i al Japó.</seg>
    </tuv>
</tu>
```

Figure 3: TMX structure of the bilingual datasets.

3. less than 200 characters in the document.
4. less than 5 segments (lines) in the document.
5. less than 20% of the segments in the document share the language identified at document level.

Cleaning further reduced the monolingual dataset size from 11 TB in the de-duplicated version to 8.4 TB. However, we believe the cleaned version is even better suited for training large language models.

4.2. Parallel Datasets

Our parallel collection is English-centric, with every language paired with English, and includes 18 language pairs with the following languages: Albanian (sq), Arabic (ar), Basque (eu), Bosnian (bs), Catalan (ca), Chinese (zh)[36], Croatian (hr), Estonian (et), Finnish (fi), Gaelic (ga), Galician (gl), Hindi (hi), Icelandic (is), Macedonian (mk), Maltese (mt), Norwegian Nynorsk (nn), Serbian (sr), and Swahili (sw). We have focused on a wide range of languages in terms of availability and language families.

The data is released in both bitext and TMX format, with the following metadata for each sentence pair: source crawl collection, Bicleaner AI score and, for each segment, the source url(s) and a hash value. An example is depicted in Figure 3.

4.2.1. Statistics

Statistics for the data release are shown in Table 2 to report about parallel segments and English-side tokens per language pair without filtering (Raw), after processing with Bicleaner AI (Filtered), and after de-duplication (De-duplicated). Raw alignments are also released to enable research on other cleaning methods or quality thresholds.

The parallel corpus contains over 96 million clean and unique sentence pairs and covers over 1.4 billion English tokens. As expected when dealing with low-resource languages, the individual corpus sizes are greatly skewed, with the top five languages accounting for 75% of the data. The average English sentence length is 14.7 whitespace-separated tokens.

The largest parallel corpora are for Finnish, followed by Arabic and Hindi. While Arabic and Hindi have a large amount of speakers (over 100 million), Finnish is far less represented on the web. The MT model used for Finnish translation is an already existing OPUS-MT model which retrieved a considerably higher number of raw aligned sentences (3.5 billion) compared to other language pairs. For Arabic and Hindi, we experimented with MT systems that were trained on data explicitly avoiding web-crawled content. Whether this approach produces a smaller, but higher quality, set of parallel candidates is still to be investigated.

Data filtering is an essential step, particularly when handling web-crawled data. We observe a 90% decrease in size when comparing the raw data with the filtered one. We also apply de-duplication, which further reduces the size by a substantial proportion, it eliminates 90% of the remaining 10%.

4.2.2. OPUS Overlaps

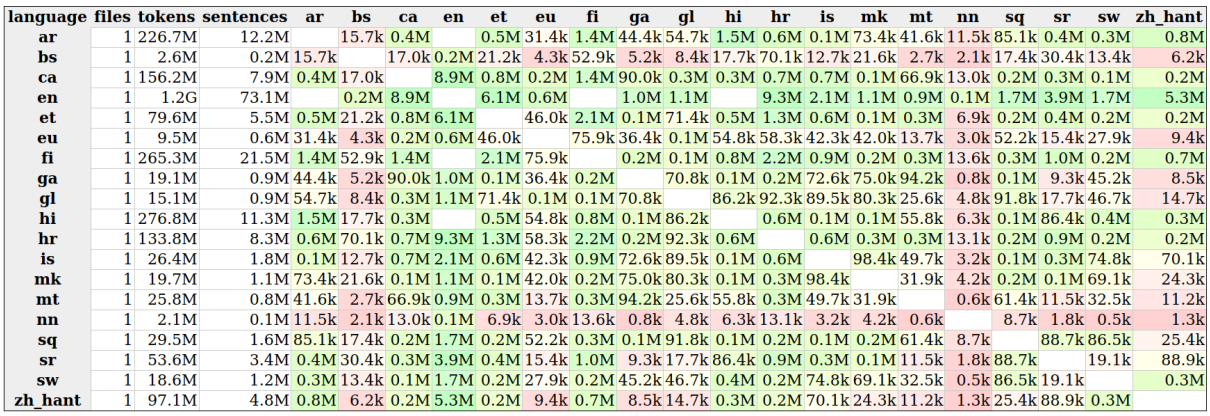
Since our parallel data is generated from the same sources as our monolingual release, which contains previously unreleased crawls, we hypothesize they also provide new information. To further investigate this issue, we have computed segment-level overlaps for each language pair with all the existing datasets in OPUS by looking at matching sentence pairs. We release detailed results for this analysis on GitHub.[37] On average, only 3.35% of our data already exists in CCMatrix and 15.72% in ParaCrawl, two of the most widely used multilingual web-crawled collections.

4.2.3. MultiHPLT

Although we plan to specifically target non-English-centric language pairs in the future, for this current release we further leverage resources by pivoting through English, creating a synthetic parallel corpus that includes all possible language pair combinations, multiHPLT. We report the statistics of those additional resources in Figure 4. It covers 171 language pairs and over 157 million parallel sentences.

[36] This release focuses on Chinese written in the traditional script.

[37] https://github.com/Helsinki-NLP/OPUS/tree/hplt2023/corpus/HPLT/v1/overlaps



language	files	tokens	sentences	ar	bs	ca	en	et	eu	fi	ga	gl	hi	hr	is	mk	mt	nn	sq	sr	sw	zh_hant
ar	1	226.7M	12.2M		15.7k	0.4M		0.5M	31.4k	1.4M	44.4k	54.7k	1.5M	0.6M	0.1M	73.4k	41.6k	11.5k	85.1k	0.4M	0.3M	0.8M
bs	1	2.6M	0.2M	15.7k		17.0k	0.2M	21.2k	4.3k	52.9k	5.2k	8.4k	17.7k	70.1k	12.7k	21.6k	2.7k	2.1k	17.4k	30.4k	13.4k	6.2k
ca	1	156.2M	7.9M	0.4M	17.0k		8.9M	0.8M	0.2M	1.4M	90.0k	0.3M	0.3M	0.7M	0.7M	0.1M	66.9k	13.0k	0.2M	0.3M	0.1M	0.2M
en	1	1.2G	73.1M		0.2M	8.9M		6.1M	0.6M		1.0M	1.1M		9.3M	2.1M	1.1M	0.9M	0.1M	1.7M	3.9M	1.7M	5.3M
et	1	79.6M	5.5M	0.5M	21.2k	0.8M	6.1M		46.0k	2.1M	0.1M	71.4k	0.5M	1.3M	0.6M	0.1M	0.3M	6.9k	0.2M	0.4M	0.2M	0.2M
eu	1	9.5M	0.6M	31.4k	4.3k	0.2M	0.6M	46.0k		75.9k	36.4k	0.1M	54.8k	58.3k	42.3k	42.0k	13.7k	3.0k	52.2k	15.4k	27.9k	9.4k
fi	1	265.3M	21.5M	1.4M	52.9k	1.4M		2.1M	75.9k		0.2M	0.1M	0.8M	2.2M	0.9M	0.2M	0.3M	13.6k	0.3M	1.0M	0.2M	0.7M
ga	1	19.1M	0.9M	44.4k	5.2k	90.0k	1.0M	0.1M	36.4k	0.2M		70.8k	0.1M	0.2M	72.6k	75.0k	94.2k	0.8k	0.1M	9.3k	45.2k	8.5k
gl	1	15.1M	0.9M	54.7k	8.4k	0.3M	1.1M	71.4k	0.1M	0.1M	70.8k		86.2k	92.3k	89.5k	80.3k	25.6k	4.8k	91.8k	17.7k	46.7k	14.7k
hi	1	276.8M	11.3M	1.5M	17.7k	0.3M		0.5M	54.8k	0.8M	0.1M	86.2k		0.6M	0.1M	0.1M	55.8k	6.3k	0.1M	86.4k	0.4M	0.3M
hr	1	133.8M	8.3M	0.6M	70.1k	0.7M	9.3M	1.3M	58.3k	2.2M	0.2M	92.3k	0.6M		0.6M	0.3M	0.3M	13.1k	0.2M	0.9M	0.2M	0.2M
is	1	26.4M	1.8M	0.1M	12.7k	0.7M	2.1M	0.6M	42.3k	0.9M	72.6k	89.5k	0.1M	0.6M		98.4k	49.7k	3.2k	0.1M	0.3M	74.8k	70.1k
mk	1	19.7M	1.1M	73.4k	21.6k	0.1M	1.1M	0.1M	42.0k	0.2M	75.0k	80.3k	0.1M	0.3M	98.4k		31.9k	4.2k	0.2M	0.1M	69.1k	24.3k
mt	1	25.8M	0.8M	41.6k	2.7k	66.9k	0.9M	0.3M	13.7k	0.3M	94.2k	25.6k	55.8k	0.3M	49.7k	31.9k		0.6k	61.4k	11.5k	32.5k	11.2k
nn	1	2.1M	0.1M	11.5k	2.1k	13.0k	0.1M	6.9k	3.0k	13.6k	0.8k	4.8k	6.3k	13.1k	3.2k	4.2k	0.6k		8.7k	1.8k	0.5k	1.3k
sq	1	29.5M	1.6M	85.1k	17.4k	0.2M	1.7M	0.2M	52.2k	0.3M	0.1M	91.8k	0.1M	0.2M	0.1M	0.2M	61.4k	8.7k		88.7k	86.5k	25.4k
sr	1	53.6M	3.4M	0.4M	30.4k	0.3M	3.9M	0.4M	15.4k	1.0M	9.3k	17.7k	86.4k	0.9M	0.3M	0.1M	11.5k	1.8k	88.7k		19.1k	88.9k
sw	1	18.6M	1.2M	0.3M	13.4k	0.1M	1.7M	0.2M	27.9k	0.2M	45.2k	46.7k	0.4M	0.2M	74.8k	69.1k	32.5k	0.5k	86.5k	19.1k		0.3M
zh_hant	1	97.1M	4.8M	0.8M	6.2k	0.2M	5.3M	0.2M	9.4k	0.7M	8.5k	14.7k	0.3M	0.2M	70.1k	24.3k	11.2k	1.3k	25.4k	88.9k	0.3M	

Figure 4: Available segments per language pair obtained via pivoting through English taken from the OPUS website. The color scale reflects the size and the counts above the diagonal refer to translation units in TMX files and below refer to aligned bitext segments, which in this case should be identical.

5. Conclusions and Future Work

In this paper, we have introduced the HPLT language resources, a new collection of monolingual and bilingual corpora leveraging data from web crawls and released under the permissive CC0 license. We have focused on curating data for medium- to low-resource languages with the hope that we will encourage the development of technology in these languages. One of our contributions is the extraction of massive multilingual text resources from Internet Archive crawls, which we have shown to offer data not found in other web-based corpora. Our data releases are, to our best knowledge, the largest *fully accessible* multilingual text corpora ever released.

While the resources presented in this paper mark a significant milestone, there are several avenues for future research. We plan to expand the language coverage and include non–English-centric language pairs; also to enrich the datasets with further metadata. Additionally, we also work to improve our tooling and, correspondingly, corpus quality. This includes deploying better language identification, tackling boilerplate identification and exploring the feasibility and benefit of performing bitexting across shards. We also seek to make better use of our HPC resources through additional automation of our data pipeline, and stability improvements for our AMD ports of MarianNMT and Megatron-DeepSpeed. In future releases of the project, we will include LLMs and MT models across our supported language set, as well as the training pipelines used to create them.

Finally, our main goal is to contribute to the NLP research field by providing massive high-quality datasets, therefore we take this opportunity to call for action and ask the community to contribute both raw data sources and processed corpora so that we can include them to our collection.

6. Environmental Considerations

Developing large-scale datasets for language modeling is an expensive task that has an environmental impact. Releasing the datasets publicly in open-source repositories directly reduces this impact, as they can be reused instead of creating from scratch. However, we believe it's important to keep track of and share how much carbon is produced when building these large datasets. Next we report our estimates in hours:

- Data download: 62K CPU
- Data processing: 72K CPU
- Monotexting: 800K CPU
- Bitexting: 2,2M CPU and 53K GPU

The total amount of hours spent would be roughly 5M CPU hours and 50K GPU hours. Note that the most expensive task is generating the parallel corpora since it involves translating all documents into English. Note also that the LUMI supercomputer uses renewable, carbon-neutral energy.[38]

7. Limitations

In this paper, we focus on the description of the construction of the first release of the HPLT language resources. We are aware that we do not provide a qualitative analysis of the datasets, and we do not train models to validate the quality of the data. While we plan to do this, these experiments are complex and expensive and a comprehensive evaluation falls out of the scope of the paper, with its main goal being to present and describe the datasets.

[38] https://lumi-supercomputer.eu/sustainable-future/

8. Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546]. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union. The authors wish to thank CSC - IT Center for Science, Finland for computational resources and support.

9. Bibliographical References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Thirteenth Language Resources and Evaluation Conference-LREC 2022*.

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518, Tampere, Finland. European Association for Machine Translation.

Marta Bañón, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, and Sergio Ortiz-Rojas. 2024. Fastspell: the Langid Magic Spell. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. In press.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages.

Andrei Z Broder. 2000. Identifying and filtering near-duplicate documents. In *Annual symposium on combinatorial pattern matching*, pages 1–10. Springer.

Miquel Esplà-Gomis, Mikel Forcada, Sergio Ortiz-Rojas, and Jorge Ferrández-Tordera. 2016. Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 685–691, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. *CoRR*, abs/1606.07947.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023. Get to know your parallel data: Performing english variety and genre classification over macocu corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, et al. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel

texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Ole Tange. 2023. Gnu parallel 20230122 ('bolsonaristas'). GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

10. Language Resource References

Adebara, Ife and Elmadany, AbdelRahim and Abdul-Mageed, Muhammad and Alcoba Inciarte, Alcides. 2023. *SERENGETI: Massively Multilingual Language Models for Africa*. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

ImaniGooghari, Ayyoob and Lin, Peiqin and Kargaran, Amir Hossein and Severini, Silvia and Jalili Sabet, Masoud and Kassner, Nora and Ma, Chunlan and Schmid, Helmut and Martins, André and Yvon, François and Schütze, Hinrich. 2023. *Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages*. Association for Computational Linguistics.

Kudugunta, Sneha and Caswell, Isaac and Zhang, Biao and Garcia, Xavier and Choquette-Choo, Christopher A and Lee, Katherine and Xin, Derrick and Kusupati, Aditya and Stella, Romi and Bapna, Ankur and others. 2023. *MADLAD-400: A Multilingual And Document-Level Large Audited Dataset*. Google.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Nguyen, Thuat and Van Nguyen, Chien and Lai, Viet Dac and Man, Hieu and Ngo, Nghia Trung and Dernoncourt, Franck and Rossi, Ryan A and Nguyen, Thien Huu. 2023. *CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages*. The University of Oregon NLP Group.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Édouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Sha Yuan and Hanyu Zhao and Zhengxiao Du and Ming Ding and Xiao Liu and Yukuo Cen and Xu Zou and Zhilin Yang and Jie Tang. 2021. *WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models*. Elsevier.

Appendix A. Source Crawl Distribution per Language

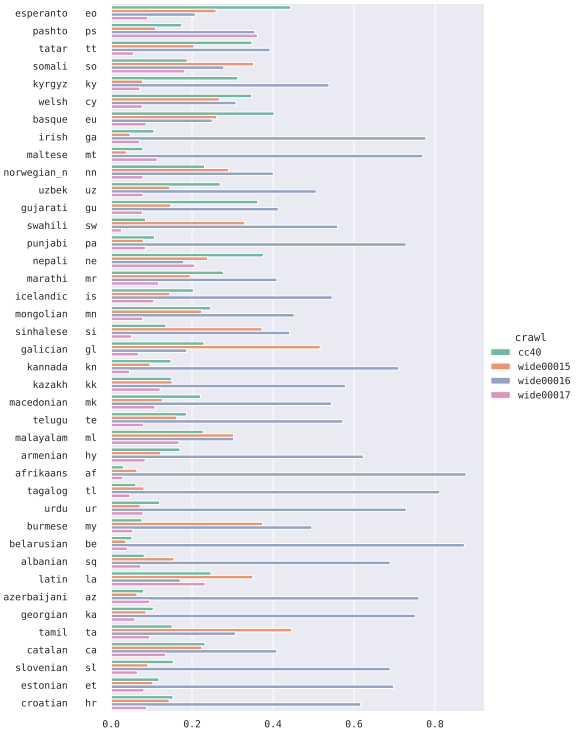


Figure 5: Proportions of text volume in bytes coming from each crawl, part 1.

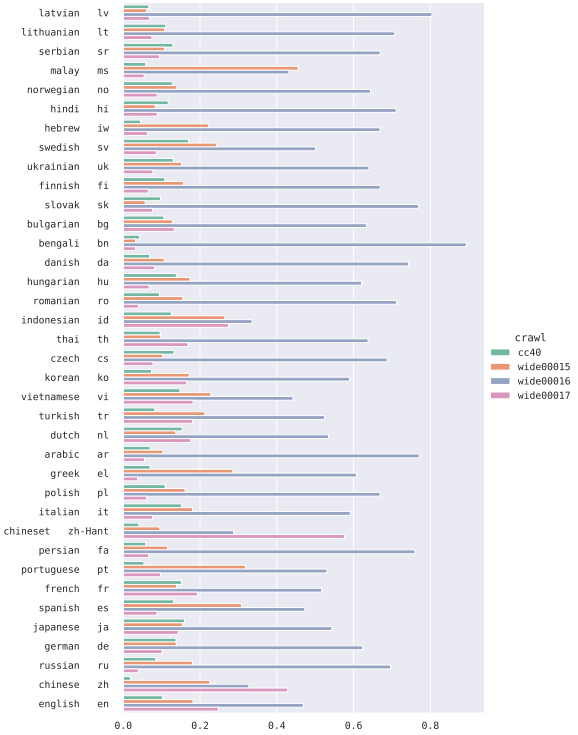


Figure 6: Proportions of text volume in bytes coming from each crawl, part 2.

Appendix B. Per-Language Monolingual Statistics

Table 3 shows the total sizes of texts in each language for the deduplicated publicly released data. The number of segments (lines), words and bytes are as reported by the Unix `wc(1)` tool, see its documentation for definitions of a line and a word. The volume of texts significantly ranges from 1.0 GB for text classified by CLD2 as Esperanto to 20.3 TB for English, while the number of documents has the minimum of 143K for Pashto and the maximum of 1.8B for English. We foresee a high percentage of documents mis-classified by CLD2 due to the huge amount of noisy data that it receives at this stage.

Language	Code	# Segments	# Words	Characters	# Bytes	# Documents
Esperanto	eo	2.29e+07	1.47e+08	9.64e+08	9.91e+08	1.77e+05
Pashto	ps	2.65e+07	1.68e+08	8.22e+08	1.32e+09	1.43e+05
Tatar	tt	2.64e+07	1.34e+08	9.65e+08	1.63e+09	1.72e+05
Welsh	cy	4.76e+07	2.44e+08	1.63e+09	1.64e+09	2.85e+05
Kyrgyz	ky	2.83e+07	1.46e+08	1.12e+09	1.96e+09	1.88e+05
Somali	so	4.41e+07	3.00e+08	1.99e+09	2.01e+09	3.75e+05
Irish	ga	1.24e+08	5.20e+08	3.38e+09	3.61e+09	9.32e+05
Norwegian	nn	1.41e+08	6.40e+08	4.36e+09	4.46e+09	7.53e+05
Basque	eu	1.46e+08	7.17e+08	5.23e+09	5.30e+09	1.01e+06
Swahili	sw	1.55e+08	9.11e+08	5.88e+09	5.94e+09	9.84e+05
Maltese	mt	1.69e+08	8.19e+08	5.69e+09	6.01e+09	4.84e+05
Gujarati	gu	8.39e+07	4.73e+08	2.95e+09	6.10e+09	4.55e+05
Uzbek	uz	9.27e+07	5.56e+08	4.38e+09	6.12e+09	6.33e+05
Punjabi	pa	1.23e+08	5.33e+08	3.06e+09	6.41e+09	8.88e+05
Galician	gl	2.22e+08	1.29e+09	8.34e+09	8.56e+09	1.79e+06
Kannada	kn	1.67e+08	5.78e+08	4.07e+09	8.71e+09	5.58e+05
Icelandic	is	3.23e+08	1.34e+09	9.16e+09	9.99e+09	1.44e+06
Tagalog	tl	2.40e+08	1.63e+09	1.15e+10	1.16e+10	1.20e+06
Sinhalese	si	1.28e+08	9.18e+08	5.82e+09	1.19e+10	5.64e+05
Macedonian	mk	2.58e+08	1.11e+09	7.31e+09	1.20e+10	1.25e+06
Mongolian	mn	1.86e+08	1.09e+09	7.55e+09	1.27e+10	1.06e+06
Marathi	mr	1.58e+08	9.12e+08	6.06e+09	1.31e+10	8.57e+05
Afrikaans	af	4.48e+08	1.87e+09	1.33e+10	1.35e+10	1.37e+06
Kazakh	kk	2.31e+08	1.01e+09	7.77e+09	1.37e+10	1.43e+06
Armenian	hy	3.55e+08	1.31e+09	9.47e+09	1.54e+10	1.36e+06
Nepali	ne	1.65e+08	1.06e+09	6.86e+09	1.56e+10	1.36e+06
Telugu	te	2.28e+08	1.06e+09	7.56e+09	1.59e+10	1.61e+06
Urdu	ur	2.68e+08	2.02e+09	1.11e+10	1.61e+10	2.23e+06
Belarusian	be	2.92e+08	1.39e+09	1.14e+10	1.93e+10	1.26e+06
Malayalam	ml	2.11e+08	1.05e+09	8.75e+09	1.94e+10	1.13e+06
Burmese	my	3.00e+08	1.25e+09	9.81e+09	1.97e+10	8.26e+05
Latin	la	5.76e+08	3.32e+09	2.40e+10	2.42e+10	4.81e+06
Georgian	ka	5.09e+08	1.68e+09	1.22e+10	2.51e+10	1.67e+06
Azerbaijani	az	7.62e+08	2.94e+09	2.22e+10	2.52e+10	3.00e+06
Albanian	sq	7.37e+08	3.78e+09	2.53e+10	2.65e+10	3.22e+06
Latvian	lv	1.48e+09	5.39e+09	3.98e+10	4.23e+10	5.12e+06
Estonian	et	1.59e+09	5.85e+09	4.48e+10	4.60e+10	5.84e+06
Slovenian	sl	1.58e+09	7.04e+09	4.79e+10	4.90e+10	5.82e+06
Catalan	ca	1.16e+09	7.88e+09	4.94e+10	5.10e+10	7.79e+06
Lithuanian	lt	1.71e+09	7.78e+09	5.40e+10	5.67e+10	7.40e+06
Tamil	ta	6.31e+08	3.87e+09	2.95e+10	6.58e+10	2.47e+06
Norwegian Bokmål	nb	3.19e+09	1.39e+10	9.21e+10	9.41e+10	1.46e+07
Slovak	sk	3.89e+09	1.39e+10	9.50e+10	1.02e+11	1.40e+07
Malay	ms	3.25e+09	1.65e+10	1.00e+11	1.02e+11	8.36e+06
Bengali	bn	2.43e+09	8.23e+09	5.51e+10	1.29e+11	5.97e+06
Finnish	fi	4.76e+09	1.69e+10	1.38e+11	1.42e+11	1.95e+07
Serbo-Croatian	hbs	4.77e+09	1.94e+10	1.32e+11	1.45e+11	1.78e+07
Hebrew	he	3.73e+09	1.55e+10	9.27e+10	1.52e+11	1.12e+07
Danish	da	4.58e+09	2.21e+10	1.53e+11	1.56e+11	2.36e+07
Hindi	hi	2.77e+09	1.37e+10	8.12e+10	1.62e+11	1.14e+07
Bulgarian	bg	3.51e+09	1.55e+10	1.04e+11	1.73e+11	1.33e+07
Swedish	sv	6.56e+09	2.83e+10	1.88e+11	1.95e+11	3.00e+07
Hungarian	hu	6.89e+09	2.65e+10	1.99e+11	2.17e+11	2.85e+07
Ukrainian	uk	3.18e+09	1.82e+10	1.34e+11	2.31e+11	1.79e+07
Romanian	ro	7.07e+09	3.28e+10	2.41e+11	2.47e+11	2.49e+07
Korean	ko	9.55e+09	2.90e+10	1.49e+11	2.80e+11	4.45e+07
Czech	cs	9.88e+09	4.10e+10	2.62e+11	2.87e+11	3.86e+07
Vietnamese	vi	9.49e+09	6.50e+10	3.17e+11	3.92e+11	4.01e+07
Thai	th	8.43e+09	2.20e+10	1.93e+11	4.05e+11	2.95e+07
Indonesian	id	9.66e+09	6.92e+10	4.81e+11	4.84e+11	4.58e+07
Turkish	tr	1.03e+10	6.49e+10	4.55e+11	4.93e+11	5.94e+07
Dutch	nl	1.65e+10	7.18e+10	5.18e+11	5.23e+11	6.66e+07
Arabic	ar	9.20e+09	5.15e+10	3.23e+11	5.27e+11	4.66e+07
Greek	el	1.13e+10	5.22e+10	3.40e+11	5.47e+11	3.06e+07
Polish	pl	1.93e+10	8.54e+10	5.95e+11	6.17e+11	8.29e+07
Persian	fa	1.34e+10	7.04e+10	3.84e+11	6.45e+11	4.23e+07
Italian	it	2.27e+10	1.14e+11	7.68e+11	7.77e+11	9.65e+07
Portuguese	pt	2.74e+10	1.32e+11	8.27e+11	8.53e+11	1.04e+08
French	fr	4.36e+10	2.14e+11	1.37e+12	1.41e+12	1.76e+08
German	de	4.35e+10	1.93e+11	1.43e+12	1.46e+12	2.26e+08
Spanish	es	4.45e+10	2.50e+11	1.56e+12	1.60e+12	2.01e+08
Japanese	ja	5.14e+10	9.14e+10	8.92e+11	1.98e+12	2.19e+08
Russian	ru	1.14e+11	4.93e+11	3.65e+12	6.02e+12	3.97e+08
Chinese	zh	1.73e+11	3.35e+11	3.35e+12	7.51e+12	1.20e+09
English	en	3.87e+11	2.86e+12	2.03e+13	2.03e+13	1.78e+09
Total		1.11e+12	5.64e+12	4.05e+13	5.01e+13	5.25e+09

Table 3: Languages in the deduplicated public data release: the number of segments (new line symbols), words (as defined by `wc(1)`), characters, bytes and documents. Ordered by size in bytes.