An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT)

Laurie Burchell[1], **Ona de Gibert**[2], **Nikolay Arefyev**[3], **Mikko Aulamo**[2], **Marta Bañón**[4], **Pinzhen Chen**[1], **Mariia Fedorova**[3], **Liane Guillou**[1], **Barry Haddow**[1], **Jan Hajič**[5], **Jindřich Helcl**[5], **Erik Henriksson**[6], **Mateusz Klimaszewski**[1], **Ville Komulainen**[6], **Andrey Kutuzov**[3], **Joona Kytöniemi**[6], **Veronika Laippala**[6], **Petter Mæhlum**[3], **Bhavitvya Malik**[1], **Farrokh Mehryary**[6], **Vladislav Mikhailov**[3], **Nikita Moghe**[7], **Amanda Myntti**[6], **Dayyán O'Brien**[1], **Stephan Oepen**[3], **Proyag Pal**[1], **Jousia Piha**[6], **Sampo Pyysalo**[6], **Gema Ramírez-Sánchez**[4], **David Samuel**[3], **Pavel Stepachev**[1], **Jörg Tiedemann**[2], **Duan Variš**[5], **Tereza Vojtěchová**[5], **Jaume Zaragoza-Bernabeu**[4]

[1]University of Edinburgh, [2]University of Helsinki, [3]University of Oslo, [4]Prompsit Language Engineering, [5]Charles University, [6]University of Turku, [7]Amazon *

Contact: https://hplt-project.org

Abstract

Training state-of-the-art large language models requires vast amounts of clean and diverse textual data. However, building suitable multilingual datasets remains a challenge. In this work, we present HPLT v2, a collection of high-quality multilingual monolingual and parallel corpora, extending prior work of the HPLT project. The monolingual portion of the data contains 8T tokens covering 193 languages, while the parallel data contains 380M sentence pairs covering 51 languages. We document the entire data pipeline and release the code to reproduce it. We provide extensive analysis of the quality and characteristics of our data. Finally, we evaluate the performance of language models and machine translation systems trained on HPLT v2, demonstrating its value.

1 Introduction

In order to train the state-of-the-art large language models (LLMs) required for modern NLP, large amounts of high-quality textual training data are essential. However, obtaining a sufficient quantity of such data is far from easy. In addition, effective NLP research requires open training data so that results can be replicated and verified.

In this paper, we introduce a new set of text corpora dubbed HPLT v2.[1] It is extracted from 4.5 petabytes (PB) of the Internet Archive (IA)[2] and Common Crawl (CC)[3] data. We build on the work of de Gibert et al. (2024) (hereafter referred to as HPLT v1.2) with an improved extraction pipeline and a much larger set of input crawls to produce the HPLT v2 collection of monolingual and parallel corpora. To our knowledge, our new corpus is the only large-scale text collection extracted from the IA, apart from HPLT v1.2. We release HPLT v2 under the permissive Creative Commons Zero (CC0) license[4] and provide the code to replicate our pipeline. Our main contributions can be summarised as:

- We release monolingual corpora covering 193 languages and containing approximately 52 trillion characters and 8 trillion tokens.
- We derive parallel corpora from our monolingual data for 50 languages paired with English, containing over 380 million sentence pairs.
- We make the tools and pipelines used to create the collection openly available.[5]
- We conduct an in-depth analysis of our data including descriptive statistics, manual inspection, and automatic register labelling.
- We demonstrate the quality of HPLT v2 by using it to train a range of high-performing language and machine translation models.

2 Related work

The increasing data demands of state-of-the-art LLMs have driven a rapid growth in both the number and the size of text corpora. We provide a summary of some well-known collections in Appendix A. Whilst LLMs trained on ostensibly English data have shown impressive multilingual capabilities (Armengol-Estapé et al., 2022), of particular relevance to this work is the growing

*Work was done prior to joining Amazon.

[1] https://hplt-project.org/datasets/v2.0

[2] https://archive.org

[3] https://commoncrawl.org

[4] https://creativecommons.org/share-your-work/public-domain/cc0/. We do not claim ownership of any of the text from which this data has been extracted.

[5] https://github.com/hplt-project/HPLT-textpipes

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17452–17485
July 27 - August 1, 2025

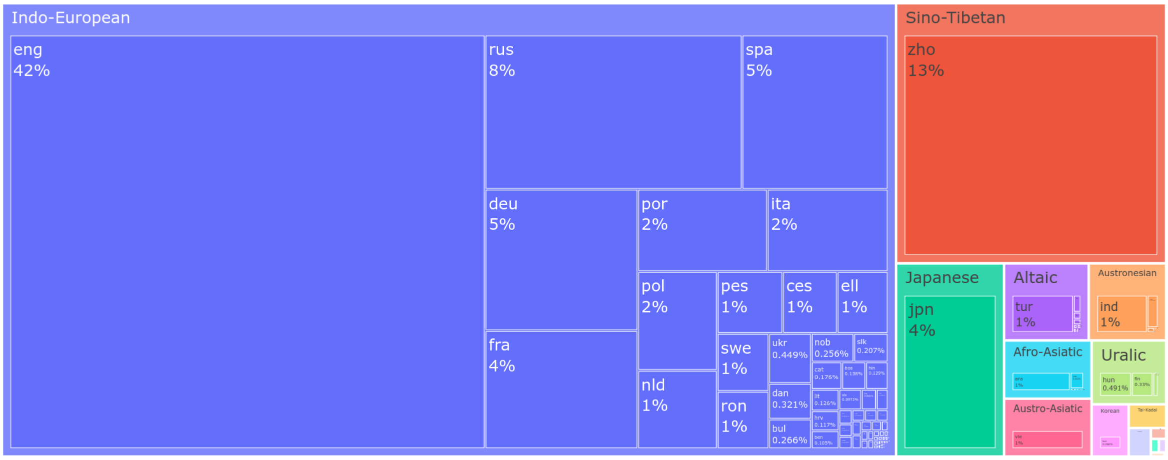

Figure 1: The distribution of documents in the HPLT v2 cleaned dataset by language family and variety. Shortened ISO 639-3 language codes are used here; see the full plots at `https://hplt-project.org/datasets/v2.0`

shift towards explicitly multilingual corpora. Compared with earlier efforts (e.g. OSCAR (Suárez et al., 2019), CC-100 (Conneau et al., 2020) and mC4 (Xue et al., 2021)), more recent multilingual datasets cover increasing numbers of languages, e.g. CulturaX (Nguyen et al., 2024a) and MADLAD-400 (Kudugunta et al., 2024). HPLT v2 continues this trend by aiming for significant coverage of a wide range of languages. We note that the majority of previous multilingual datasets are sourced from CC, whereas much of HPLT v2 is composed of IA crawls. This means that HPLT v2 can be used in conjunction with these existing datasets as a complementary source.

Producing large-scale datasets by crawling the Web is helpful for scale, but raises questions around dataset quality such as the prevalence of boilerplate, explicit material or non-linguistic content (Kreutzer et al., 2022). One way to tackle low-quality data is through human audit and curation (e.g. ROOTS (Laurençon et al., 2022), Glot500-c (Imani et al., 2023), Serengeti (Adebara et al., 2023) and the MaLA Corpus (Ji et al., 2024)). However, such an approach is difficult to scale. Instead, we ensure the quality of HPLT v2 through a robust dataset construction pipeline (Section 4) and by verifying our data through extensive analysis and downstream evaluation (Sections 5 and 6).

In addition to large-scale monolingual data in multiple languages, HPLT v2 contains high-quality parallel data. Whilst causal language models (CLMs) with decoder-only architectures rely primarily on monolingual data, recent studies have shown that incorporating parallel data during the pretraining stage significantly boosts multilingual, cross-lingual and machine translation (MT) performance for such models (Kale et al., 2021; Briakou et al., 2023; Alves et al., 2024). Because of this, we expect that there is still significant demand for parallel data.

This work is a direct successor of our HPLT v1.2 dataset introduced in (de Gibert et al., 2024). Compared to HPLT v1.2, we feature more data (21 billion vs. 5 billion documents) using an improved pipeline (Section 4), resulting in a significantly larger dataset (52 trillion characters compared to 42 trillion). The HPLT v2 collections are of higher quality than those in HPLT v1.2, as shown through comparative analysis and evaluation (Sections 5 and 6). For clarity, we list the main differences between HPLT v1.2 and HPLT v2 here:

- The size of the source web collections is 2.5x larger, totalling 4.5 petabytes of compressed web data.
- The text extraction pipeline uses `Trafilatura` rather than `warc2text`, which results in more efficient boilerplate removal.
- Language identification uses a modified version of `OpenLID` rather than `CLD2`, increasing coverage from 75 to 193 language varieties.
- Documents are annotated with their compliance to the `robots.txt` files of the corresponding websites, which can be used to filter out documents explicitly forbidden for crawling by website owners. The cleaned

variant of the HPLT v2 dataset contains only `robots.txt` compliant documents.
- Deduplication is done at collection level rather than globally.
- Documents are annotated for PII information.
- Document quality scores computed with `web-docs-scorer` rather than segment-level language model based scores
- Simpler and more interpretable filtering and cleaning criteria.

3 Dataset description

In this section, we describe the HPLT v2 collection of monolingual and parallel corpora, before explaining how it was constructed in Section 4.

3.1 Monolingual datasets

The monolingual portion of HPLT v2 covers 193 language varieties[6] and is published in two variants: 'deduplicated' (21 terabytes) and 'cleaned' (15 terabytes). In the latter variant, the documents filtered by our cleaning heuristics (see Section 4.2) are excluded. For training LLMs, we recommend using the cleaned variant, but we also publish the datasets before cleaning ('deduplicated') so that it is possible to apply custom cleaning pipelines to the HPLT v2 data. In total, the deduplicated monolingual HPLT v2 datasets contain approximately 7.6 trillion white-space separated tokens and 52 trillion characters, extracted from 21 billion documents. HPLT v2 is published in the JSONL format, with one document per line.

Figure 1 shows the distribution of documents in the cleaned monolingual data by language families and language variety. Indo-European languages, and especially English, make up the majority of the data. Unfortunately, this is the reality of current web crawls; increasing the amount of data available for other languages is not an easy task and is important future work. Appendix B gives a full breakdown of the statistics of the monolingual data.

3.2 Parallel datasets

We extract parallel data from the monolingual HPLT v2 to cover 50 languages paired with English. We aimed for a diverse range of language varieties and scripts in the low to medium resource range (listed in Table 5). We align these to English since this configuration has the highest potential for finding high-quality parallel data. We release our data in both XML and bitext format.[7]

	Raw		Filtered	
	Pairs	**Eng. words**	**Pairs**	**Eng. words**
Total	1277M	16849M	380M	6780M
Median	11M	170M	4M	80M

Table 1: Counts in millions (M) of sentence pairs and English words in the parallel HPLT v2 data before filtering (Raw) and after filtering and deduplication (Filtered), both in total and the median across all languages.

Table 1 gives the number of sentence pairs and English words per language prior to filtering (Raw) and after processing (Filtered). We provide both the total over the entire dataset and the median count by language variety. Our results show that the deduplicated HPLT v2 parallel corpora have a 70% reduction in sentence pairs compared to the raw data. The final dataset contains over 380 million sentence pairs, with the English side of the dataset containing over 6 billion words. The median number of sentence pairs by language variety is 4 million, but individual sizes vary greatly by language: the smallest, Sinhala, contains around 273 thousand pairs, whereas the largest, Finnish, contains over 29 million pairs. We give full statistics for each included language variety in Table 5 in the appendix.

We assume the large number of Finnish sentence pairs is due to the pipeline's bias toward European languages. In contrast, languages such as Japanese and Korean, which we would expect to have larger corpora, may have lower counts because of lower-quality monolingual data and limited support in key pipeline components such as sentence splitting and tokenization. This results in reduced yields during data cleaning and filtering for non-European languages written in non-Latin scripts.

DocHPLT v2 We also introduce context, by providing the documents that our bitext data comes from. These documents are annotated with both sentence and paragraph alignment. Overall, we provide 74,078,581 aligned documents covering all 51 languages. This ranges from 38,199,407 documents in English to 20,448 documents in Xhosa.[8]

MultiHPLT v2 We leverage the English-centric HPLT v2 parallel resources to further create a multi-

[6]Language varieties are labelled with an ISO 639-3 code denoting the variety plus an ISO 15924 four-letter code denoting the script, separated by an underscore: e.g., `gla_Latn`.

[7]https://opus.nlpl.eu/HPLT/corpus/version/HPLT

[8]https://opus.nlpl.eu/legacy/DocHPLT.php

way parallel corpus, obtained by pivoting via English sentences. This corpus includes 1275 language pairs and contains over 16.7 billion parallel sentences.[9]

A similar pivoting is also done for DocHPLT and we create the additional bitexts through sentences aligned to the same English document. However, anchoring sentences in aligned documents creates a different subset across the pivoted language pairs.

4 Dataset construction

In the following section, we explain the dataset construction pipeline for HPLT v2. We first extract text from web crawls via HTML (Section 4.1), deduplicate and clean this monolingual text (Section 4.2), and finally extract and process the parallel data (Section 4.3). Figure 2 provides a high-level overview of the pipeline.

4.1 Text extraction from web crawls

Sources In total, we ingest 4.5 PB of web crawl data to build HPLT v2. 3.7 PB is sourced from IA from crawls conducted mostly between 2012 and 2020, with the remaining 0.8 PB coming from CC. We use CC crawls conducted mostly between 2014 and 2022. A detailed description of the crawls we use is in Appendix C.

Extracting HTML Both IA and CC crawls are provided in the Web ARChive (WARC) format[10] which stores HTTP requests and responses between a web crawler and web servers. We use the `warc2text` tool[11] to extract HTML and related metadata from these WARC files. It selects relevant WARC records containing HTML pages, removes documents from a list of known trash websites,[12] and finally saves the results in the ZSTD-compressed JSONL format. The extracted metadata includes document URLs, paths to the original WARC files and record positions inside, content types and timestamps. Additionally, WARC records with URLs ending with "`robots.txt`" are stored for later use in filtering.

Extracting text This stage of the pipeline extracts the main textual content from HTML

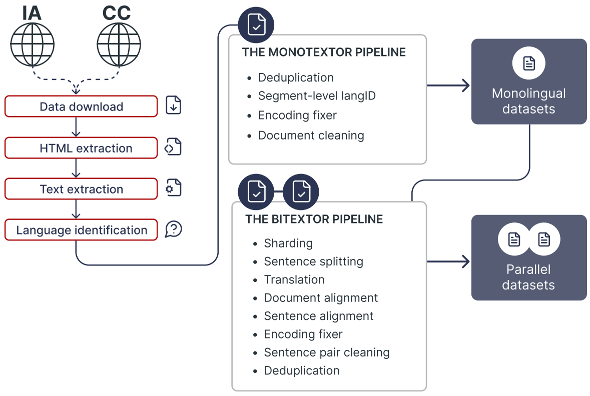

Figure 2: Overview of the data acquisition and processing pipeline for HPLT v2.

pages and groups it into language-specific subsets. It first parses the HTML pages into a tree representation. Next, it removes likely machine-translated texts by searching for indicative HTML tags and attributes. It then removes boilerplate (i.e. parts of a web page that do not contribute to its main content) using Trafilatura 1.8.0 (Barbaresi, 2021). Following hyperparameter experimentation, we set `include_comments=False`, `include_tables=False`, `no_fallback=False` and `MIN_EXTRACTED_SIZE=0`, with all other hyperparameters set to their defaults. We chose not to use fallback to Trafilatura's simple extraction baseline since it leaves most boilerplate intact, and we preferred sacrificing some documents but avoiding extra boilerplate in HPLT v2. Finally, we predict the language of the text using a modified version of the OpenLID model[13] (Burchell et al., 2023; Burchell, 2024), where the Arabic dialects are combined under one macrolanguage label and the model training data has undergone improved pre-processing. These changes are intended to improve classification reliability. After text extraction, the dataset size reduces to 62 TB, 15 times smaller than the HTML data and 75 times smaller than the original web crawls.

4.2 Monolingual text processing

Following text extraction, we proceed to monolingual text cleaning in which we apply various criteria to select the cleanest documents. With the exception of fixing encoding, we do not alter the text in this process.

We first discard all documents for which the predicted probability of the language label is < 0.5.

[9] https://opus.nlpl.eu/MultiHPLT/corpus/version/MultiHPLT

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

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

[12] Mostly containing auto-generated lists of phone numbers, addresses, etc.: https://github.com/paracrawl/cirrus-scripts/blob/master/url-filter-list.annotated

[13] https://data.statmt.org/lid/lid193_merged_arabics.bin

We then perform crawl-level deduplication with a MinHash index (Broder et al., 1998), using 240 hashes and a Jaccard similarity threshold of 0.8. We keep one document from each computed disjoint-set (Galler and Fischer, 1964), thus removing near-duplicates within each crawl.

To respect `robots.txt`[14] rules specified by each domain, we use the extracted `robots.txt` records to identify patterns disallowing the crawlers we use.[15] We use the `fst`[16] tool to create a compressed index of URLs to exclude and use it to remove documents originating from these URLs.

We then use a range of heuristics to discard low-quality documents. We calculate a document quality score using Web Docs Scorer (WDS),[17] discarding documents with a score < 5. We remove any documents where the length of the document is < 500 characters, or where the average number of words per segment is < 5 (< 10 characters for Japanese, Chinese or Korean). We also filter documents where the URL is in the UT1 adult list.[18]

Finally, we enrich documents with additional metadata. We add a unique identifier for the document hash derived from the WARC file name, the URL and the timestamp. We also carry out segment-level language identification (LID) using the Rust port[19] of HeLI-OTS (Jauhiainen et al., 2022), trained on the OpenLID dataset. Finally, we add the Unicode character offsets of any personally identifiable information found by the PII tool.[20]

Although the CC crawls are less than 20% of the input data, they are the source of about 60% of the final text. This is likely because CC focuses on textual content whereas IA includes much multimedia content, resulting in 4-8x lower yields in general. However, for some languages (e.g. Chinese, Persian, and a few smaller languages), IA provides more texts than CC. Appendix D presents a detailed study of the contributions of different source crawls to the final dataset.

[14] `https://www.robotstxt.org/`

[15] `*`, `CCBot`, `ia-archiver`

[16] `https://burntsushi.net/transducers/`

[17] `https://github.com/pablop16n/web-docs-scorer/`

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

[19] `https://github.com/ZJaume/heliport`

[20] `https://github.com/mmanteli/multilingual-PII-tool`

4.3 Parallel data extraction

Our parallel data extraction pipeline is adapted from Bitextor.[21] We make the following changes to increase the quality of the final dataset:

- Input data comes from cleaned monolingual HPLT v2 rather than WARCs.
- We use Loomchild, a SRX-based sentence splitter (Miłkowski and Lipski, 2011), to cover more languages.
- During sentence splitting, paragraph and sentence identifiers are added as persistent metadata through the pipeline.
- Minimal length rule and fluency filtering in bicleaner-hardrules are disabled as this duplicates other processing steps.
- Bicleaner AI (Zaragoza-Bernabeu et al., 2022) uses a multilingual model able to handle unseen language pairs during training.
- Document-level output from the document-matching step is collected to allow the creation of document-level parallel data.

To avoid the possible introduction of new bugs in the pipeline, given that many of the steps in it are made for 2-letter language codes, we convert 3-letter language codes to 2-letter before processing.

5 Data analysis

In this section, we present our analysis of the HPLT v2 data based on indirect quality indicators, manual inspection, and register labels.

5.1 Indirect quality indicators

We consider two types of indirect quality indicators: descriptive statistics and website domains.

Descriptive statistics We calculate descriptive statistics for HPLT v2 using the HPLT Analytics tool.[22] We compare to the cleaned HPLT v1.2 dataset (our previous release).

For the monolingual data, there are far more unique segments (22.2% of HPLT v1.2 vs. 40.9% of HPLT v2) but far fewer documents longer than 25 segments (90.8% vs. 23.2%). Similarly, the proportion of short segments is reduced (39.6% vs. 13.3%). These changes can be attributed to the use of Trafilatura and WDS. More segments match the document language (58.6% vs. 81.5%), driven by improvements in LID accuracy and a more aggressive document filtering strategy. Finally, we iden-

[21] `https://github.com/bitextor/bitextor`

[22] `https://github.com/hplt-project/data-analytics-tool`

tify frequent n-grams and find that a substantial amount of textual boilerplate remains, particularly from Wikipedia and blogging platforms.

Regarding parallel data, we find that the number of source and target tokens per language pair is much higher in HPLT v2 (by 47% and 49% on average) than in HPLT v1.2. Furthermore, 80% of the sentence pairs have a translation likelihood score from 0.8 to 1 (as computed by Bicleaner AI) which attests to their high quality. The frequent n-grams in the parallel datasets are similar among all languages: larger datasets tend to focus on hotels and legal notices, whereas the smaller datasets exhibit more variety and the frequent n-grams in these datasets reflect local content likely from news websites, e.g. political figures and place names. Appendix D contains further examples.

Domains We explore the website domain names and geographic top-level domains (TLDs) present in the data in order to understand its origins better.

We find different patterns of website domain names in the corpora depending on the size of the language dataset. Languages with more data available contain a diverse range of website domain names in the monolingual data but more travel-related webpages in the parallel data. However, smaller language datasets tend to contain more Wikipedia and religious content in both the monolingual and parallel data. Appendix F contains further information about common domains.

Whilst most of the TLDs in our dataset are general purpose (e.g. `.com`, `.org`), we found that the most common geographic TLDs in the monolingual language corpora were usually from the country with the most speakers. This gave us confidence in the reliability of the text. We found the proportion of geographic TLDs from an indicative country was highest for those in Europe, whereas the datasets for many languages primarily spoken in Africa mostly consisted of general-purpose TLDs. The parallel data exhibits more diversity in TLDs than the monolingual data. For example, `.eu` is much more frequent, appearing in the top-10 TLDs of all mid-size and large parallel datasets of nearly all European languages. A more detailed discussion of our observations is in Appendix G.

5.2 Manual data inspection

To assess human-perceived quality, we manually inspected a random sample of documents from the cleaned monolingual datasets in 22 languages. Specifically, for each language spoken by the authors, we sampled 50 random documents extracted from each of the four groups of crawls: the older CC/IA crawls from 2012–2014, and the newer CC/IA crawls from 2017–2020. The main goal of this stratification was to compare the quality of texts we get depending on the crawl source and age, and select the most promising crawls for the next release of our datasets.

We asked participants to annotate any documents which look like pornographic content, look unnatural, and/or are not in the target language. Appendix H describes the inspection procedure and the results. Overall, for most languages, both the proportions of pornographic content and texts not in the target language are around 0–3%, with no significant difference between groups of crawls. Asturian, Scottish Gaelic and Norwegian Nynorsk are notable exceptions, with 31, 11 and 7 percent of texts not in the target language respectively. The proportion of unnatural texts is around 10% on average and up to 30% for some languages, leaving space for improvements. We also observe that the probability of getting an unnatural text from the newer CC crawls is roughly half of that of the other three inspected groups of crawls. This is probably related to the introduction of harmonic centrality ranking for domain prioritization in the CC crawler queue since 2017 (Nagel, 2023), which is stated to be more efficient in avoiding spam compared to the previously used techniques.

5.3 Register labels

As noted in Section 5.1, web crawls cover a vast range of different kinds of documents from various sources. We use automatic register (or genre) classification to create metadata about this variation, allowing users to make informed decisions when sampling from the data.

We use the multilingual register classifier described in Henriksson et al. (2024) to label the entire monolingual HPLT v2. This classifier covers 16 languages and is based on an XLM-R Large model (Conneau et al., 2020), fine-tuned on a multilingual web corpus manually annotated with register information. The classifier employs a hierarchical taxonomy with 25 register classes organized into 9 main categories (listed in Table 2). This label scheme is specifically designed for the linguistic characteristics of web texts and is therefore well-suited for our dataset.

The system achieves a mean micro F1 score of

Register	Percentage
How-to Interactive (HI)	1.8 %
Interactive Discussion (ID)	6.5 %
Informative Description (IN)	27.1 %
Informative Persuasion (IP)	10.8 %
Lyrical (LY)	0.5 %
Machine Translated (MT)	3.3 %
Narrative (NA)	18.1 %
Opinion (OP)	5.4 %
Spoken (SP)	0.2 %
Multiple labels	23.6 %
No label	2.5 %

Table 2: Register label distribution in HPLT v2 English dataset for classification threshold 0.4. See Henriksson et al. (2024) for the full scheme and explanation of the contents of the classes.

77% on the 5 languages used during fine-tuning. It also demonstrates good performance for 11 unseen languages, with a mean micro F1 score of 66%. These results allow us to extend register labelling to a broad range of languages, though we limit predictions to languages within the 100 languages covered by XLM-RoBERTa.

We provide the classification certainty as well as the label, so that the threshold can be optimized by use case. Table 2 presents the distribution for register labels in our English data for a classification threshold of 0.4. Further work could use our derived labels to improve dataset quality, by e.g. filtering out MT content.

6 Empirical evaluation

In this section, we describe our empirical evaluation of the quality of the HPLT v2 monolingual corpora. We conduct this evaluation by employing the datasets as training material for several natural language processing (NLP) models.

6.1 Basic linguistic tasks and MLMs

We train masked language models (MLMs) on 52 different languages from the HPLT v2 datasets, choosing those with available benchmarks.[23] We use LTG-BERT (Samuel et al., 2023) to allow comparison with HPLT v1.2. We give full details about LTG-BERT in Appendix J.

We evaluate the trained MLMs on part-of-speech tagging, lemmatization and dependency parsing using the Universal Dependencies (UD) treebanks (de Marneffe et al., 2021), as well as named entity recognition (NER) using WikiAnn datasets (Pan et al., 2017). We compare to mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) models as multilingual baselines, and to HPLT v1.2 BERT models[24] as monolingual baselines. The performance is measured using the official CoNLL 2018 evaluation code (Zeman et al., 2018) for the UD tasks, and `seqeval` (Nakayama, 2018) balanced F1 score for the NER task.

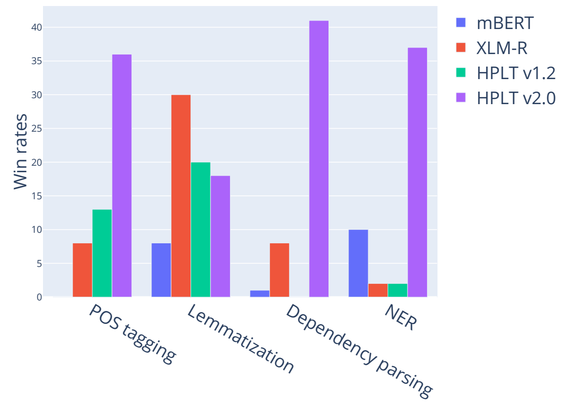

Figure 3: Win rates for MLMs at part-of-speech tagging, lemmatisation, dependency parsing, and named entity recognition.

Figure 3 shows the win rates achieved by the models for the four tasks ('win rate' here is the number of languages on which a given model outperforms other models). Models trained on the HPLT v2 datasets show a considerably higher win rate compared to the baselines in all the tasks except lemmatization, where XLM-R and HPLT v1.2 yield competitive results. However, we note that the difference between XLM-R, HPLT v1.2 and HPLT v2 on the lemmatization task is less than 1% of accuracy, meaning that no model significantly outperforms any other. Detailed scores by language and task are to be found in Table 15. We make the HPLT v2 BERT models with intermediate checkpoints publicly available.[25]

6.2 NLU tasks and large generative LMs

Pretraining generative language models (LMs) and evaluating their downstream performance on advanced natural language understanding (NLU) tasks is an established way to measure and compare training data quality (Gao et al., 2020; Penedo et al., 2023; Longpre et al., 2024). Following Penedo et al. (2024a), we compare various large web-crawled pretraining corpora using this method

[23] https://github.com/hplt-project/HPLT-WP4

[24] https://hf.co/collections/HPLT/hplt-bert-models-6625a8f3e0f8ed1c9a4fa96d

[25] https://hf.co/collections/HPLT/hplt-20-bert-models-67ba52ae96b1fb8aae673493

for one high-resource and one low-resource language: English and Norwegian. We train 1.7B decoder-only LMs using 100B/30B tokens sampled from the English/Norwegian parts of our HPLT v2 dataset respectively. We compare our English and Norwegian models with models trained on same-sized samples of HPLT v1.2 (de Gibert et al., 2024) and FineWeb (Penedo et al., 2024a), and additionally compare our Norwegian models with FineWeb-2 (Penedo et al., 2024b), CulturaX (Nguyen et al., 2024b), and mC4 (Xue et al., 2021). We replicate the design by Penedo et al. (2024a) and train the models with a fixed pretraining setup except for the pretraining corpus (English: four corpora; Norwegian: five corpora). We provide full details on pretraining and evaluation in Appendix I and describe our key results below.

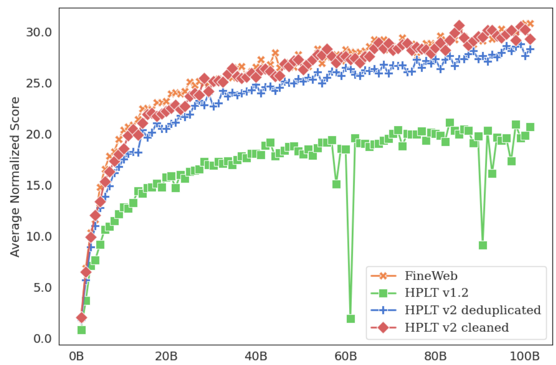

Figure 4: Performance comparison of the trained generative LMs on English.

English Average results over the English benchmarks are presented in Figure 4. Our models trained on the cleaned HPLT v2 datasets reach similar performance to the models trained on FineWeb data in downstream tasks, and considerably outperform the models trained on HPLT v1.2 and on the deduplicated subset of HPLT v2. This implies that our cleaning approach has successfully improved the data quality with respect to these benchmarks.

Norwegian Average normalized scores over the Norwegian tasks are shown in Figure 5. We observe that the Norwegian models trained on FineWeb, CulturaX, and mC4 perform on par with HPLT v2 and outperform those trained on HPLT v1.2. Performance gains start to level off after 16B tokens, with the FineWeb and HPLT v2 scores being more stable during pretraining. This suggests that CulturaX, FineWeb, and HPLT v2 are more effective corpora for Norwegian, and their

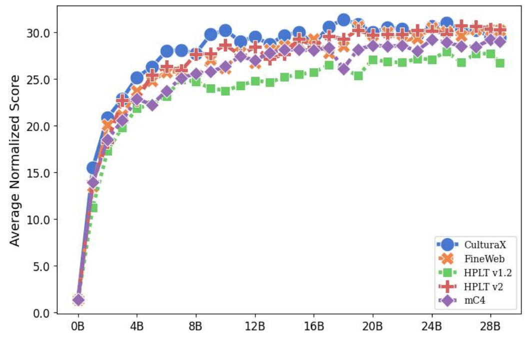

Figure 5: Performance comparison of the trained generative LMs on Norwegian.

mixtures potentially provide further benefits.

6.3 Machine translation tasks

We extrinsically evaluate the quality of the HPLT v2 parallel data by measuring the performance of MT models trained on it. This section considers two settings:

- We compare HPLT v2 to its predecessor, HPLT v1.2, where we measure their performance in 10 overlapping languages.
- We assess HPLT v2 as a complementary dataset to existing MT resources. We compare models in three data conditions: 1) solely on HPLT v2 parallel data; 2) on the data from Tatoeba Challenge (which includes most of the OPUS collection for each direction Tiedemann, 2012, 2020); and 3) on a combination of the two.

We carry out training for each individual language pair (to and from English) using the Transformer-base architecture (Vaswani et al., 2017) and the Marian NMT toolkit (Junczys-Dowmunt et al., 2018). Data processing and training are streamlined with OpusPocus[26] following the configuration of Arefyev et al. (2024). We evaluate all models on the FLORES-200 benchmark (NLLB Team et al., 2024) using BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET-22-DA (Rei et al., 2022). We use sacrebleu's implementation of BLEU[27] and chrF++[28] with signatures footnoted (Post, 2018). Full results for each metric and each direction are given in Table 16 and Table 17 in Appendix K.

We report average BLEU and COMET scores in Table 3. It is worth noting that we only average

[26] `https://github.com/hplt-project/OpusPocus`

[27] `nrefs:1|case:mixed|eff:no|smooth:exp|version:2.5.1`, and where applicable, `tok:ja-mecab`, `tok:ko-mecab`, or `tok:13a`

[28] `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`

	xx-en		en-xx	
	BLEU	COMET	BLEU	COMET
HPLT v1.2	28.5	0.7943	24.4	0.7623
HPLT v2	**32.7**	**0.8343**	**27.9**	**0.8137**
HPLT v2	28.7	0.8144	23.4	0.7941
OPUS	29.6	0.8142	23.4	0.8074
HPLT v2+OPUS	**30.5**	**0.8237**	**24.2**	**0.8083**

Table 3: Average scores for the HPLT v1.2 and v2 comparison (top) and HPLT v2 as a complimentary resource to OPUS (bottom). Only numbers that are available for all models in a comparison are averaged.

numbers for translation directions that are covered by all models in a comparison. The top half features a clear advantage of HPLT v2 over HPLT v1.2, reflecting the effect of our improved data extraction pipeline discussed in Section 4. Results in the bottom of Table 3 shows that HPLT v2 and OPUS perform comparably. However, combining the two datasets leads to improvements in both BLEU and COMET. This indicates that HPLT v2 contains non-overlapping content relative to existing OPUS corpora, making it a valuable complementary resource for MT.

7 Conclusions and future work

This paper introduced the HPLT v2 dataset, a large-scale multilingual collection of openly-available monolingual and parallel web-crawled data. Building on our previous HPLT v1 effort, we focused on improving the quality of data available for a wider range of languages, and we make our data processing pipelines publicly available for easy reuse. We presented extensive data analysis as well as intrinsic and extrinsic evaluation, demonstrating the value of HPLT v2 for various NLP tasks.

Further work will focus on expanding language coverage and data quality, particularly for underserved languages. We also plan to conduct more experiments on the document-level aligned parallel corpus.

Limitations

Like many large-scale corpora, the majority of the data in HPLT v2 is in Indo-European languages, especially English, and the parallel data is English-centric. To an extent, this is a result of the dominance of these languages in the source web-crawl data. In addition, the evaluation in the paper only covers a subset of the languages in HPLT v2 due to a lack of resources for all languages present. We hope that the data we release in multiple underserved languages will be used to improve language technologies for more communities.

Whilst we focus on improving the HPLT v2 data processing pipeline, there are still residual errors in the final dataset in LID, boilerplate removal (particularly Wiki* boilerplate) and other cleaning steps. We make the code for our pipeline available to facilitate its evaluation and improvement. We note that there is only limited removal of machine-generated content in HPLT v2 (i.e., content generated by technologies like MT and LLMs), as detecting such content remains a difficult task (Yang et al., 2024).

It is possible that some of the test data we use for evaluation is contained within HPLT v2 (for example, the Wikipedia-based test set for named entity recognition). Nevertheless, we believe that the results reported in Section 6 are still indicative of the quality of HPLT v2, since the large-scale datasets we compare against are likely to have similar contamination issues.

During the evaluation, we discovered that the punctuation for Chinese languages (and probably Korean and Japanese) in HPLT v2 had been normalised incorrectly to its Latin equivalent, causing a drop in measured performance for languages in this script. We will fix this in the next iteration of the HPLT pipeline.

Ethical considerations

We source our data from web crawls and since Internet text is largely unregulated, our final dataset may contain harmful content or amplify existing biases, despite the extensive filtering applied to mitigate these issues. One notable bias is the over-representation of religious content in smaller language corpora, which could lead to models trained on this data being biased towards this particular domain.

Another pressing ethical consideration is the significant environmental impact of producing large-scale datasets. We mitigate this impact by making the data openly available in multiple formats, limiting the need to reproduce the processing pipeline.

We report the estimated CPU and GPU cost in hours for our work to allow for more informed decision-making in future research efforts:

- WARC to HTML extraction: 250K CPU
- HTML to text extraction: 1.7M CPU
- Monolingual data processing: 600K CPU
- Parallel data cleaning and deduplication:

1.8M CPU and 23K GPU

- Register labels classification: 36.7K GPU
- MLM experiments: 1.8K CPU and 4K GPUs
- Generative LMs training and evaluation: 21.5K CPU and 43K GPU
- MT models training and evaluation: 20K GPU

The total amount of hours spent would be roughly 4.4M CPU hours and 106K GPU hours. The most expensive task is the evaluation of our data through generative model training. We mitigate the environmental impact of our work by using one of the most eco-efficient HPC clusters in the world (LUMI) to carry out much of our computation. The LUMI supercomputer uses renewable, carbon-neutral energy.[29]

Acknowledgements

We acknowledge Tsz Kin Lam, Otto Tarkka, and Maryam Teimouribadelehdareh for their annotation work; we thank Jelmer van der Linde for his help with Go.

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. The project has also been supported by the Czech MEYS project No. CZ.02.01.01/00/23_025/0008691 and Research Infrastructure project LM2023062.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. *arXiv preprint*.

Nikolay Arefyev, Mikko Aulamo, Pinzhen Chen, Ona De Gibert Bonet, Barry Haddow, Jindřich Helcl, Bhavitvya Malik, Gema Ramírez-Sánchez, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, and Jaume Zaragoza-Bernabeu. 2024. HPLT's first release of data and models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 53–54, Sheffield, UK. European Association for Machine Translation (EAMT).

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. 1998. Min-wise independent permutations (extended abstract). In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, page 327336, New York, NY, USA. Association for Computing Machinery.

Laurie Burchell. 2024. *Improving natural language processing for under-served languages through increased training data diversity*. Ph.D. thesis, University of Edinburgh.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question

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

answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

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, Online. Association for Computational Linguistics.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Bernard A. Galler and Michael J. Fischer. 1964. An improved equivalence algorithm. *Commun. ACM*, 7:301–303.

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*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. 2024. Automatic register identification for the open web using multilingual deep learning. *Preprint*, arXiv:2406.19892.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. NorQuAD: Norwegian question answering dataset. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. EMMA-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmT5 - is parallel data still relevant for pre-training massively multilingual language models? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 683–691, Online. Association for Computational Linguistics.

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.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Hynek Kydlíček, Guilherme Penedo, Clémentine Fourier, Nathan Habib, and Thomas Wolf. 2024. FineTasks: Finding signal in a haystack of 200+ multilingual tasks.

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.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Vladislav Mikhailov, Tita Enstad, David Samuel, Hans Christian Farsethås, Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2025a. NorEval: A Norwegian Language Understanding and Generation Evaluation Benchmark. *arXiv preprint arXiv:2504.07749*.

Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025b. A collection of question answering datasets for Norwegian. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 397–407, Tallinn, Estonia. University of Tartu Library.

Marcin Miłkowski and Jarosław Lipski. 2011. Using srx standard for sentence segmentation. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 172–182, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. *arXiv preprint arXiv:2406.07545*.

Sebastian Nagel. 2023. Common Crawl: Data collection and use cases for NLP. *HPLT & NLPL Winter School on Large-Scale Language Modeling and Neural Machine Translation with Web Data, February*, 6.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024a. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024b. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

NLLB Team et al. 2024. No language left behind: Scaling human-centered machine translation. *Nature*, 630(8018):841.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The FineWeb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíek, Vinko Sabolec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. FineWeb2: A sparkling update with 1000s of languages.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Samuel. 2023. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 221–237, Singapore. Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *Preprint*, arXiv:2201.11990.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

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, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. 2024. A survey on detection of LLMs-generated content. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9786–9805, Miami, Florida, USA. Association for Computational Linguistics.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Comparison of multilingual collections

Dataset	Size (TB)	Tokens (T)	Langs	% English	Source
English only					
The Pile (Gao et al., 2020)	0.8	0.39	1	100	various (c.f Section 2)
C4.en (Raffel et al., 2020; Dodge et al., 2021)	0.3	0.16	1	100	Common Crawl
RefinedWeb (Penedo et al., 2023)	2.8	0.6	1	100	Common Crawl
Dolma-Web (Soldaini et al., 2024)	-	2.28	1	100	Common Crawl
FineWeb (Penedo et al., 2024a)	-	15	1	100	Common Crawl
Multilingual					
OSCAR-23.01 (Suárez et al., 2019)	-	1.1	153	48.43	Common Crawl
CC-100 (Conneau et al., 2020)	2.39	0.3	100	18.84	Common Crawl
mC4 (Xue et al., 2021)	-	6.3	101	5.67	Common Crawl
ROOTS (Laurençon et al., 2022)	1.6	0.4	46	30.03	BigScience Catalogue Data, Common Crawl, OSCAR
Glot500-c (Imani et al., 2023)	0.6	-	511	*2.16	various (c.f. Appendix C)
Serengeti (Adebara et al., 2023)	0.042	-	517	-	various (c.f. Appendix C)
CulturaX (Nguyen et al., 2024a)	27	6.3	167	45.13	OSCAR, mC4
MADLAD-400-clean (Kudugunta et al., 2024)	-	2.6	419	50	Common Crawl
MaLA (Ji et al., 2024)	-	0.074	939	4	various (c.f. Section 2.1.4)
monoHPLT v1.2-dedup (de Gibert et al., 2024)	11	5.6	75	41	Common Crawl, Internet Archive
HPLT v2 (monolingual, deduplicated)	21	7.6	193	44	Common Crawl, Internet Archive

Table 4: Comparison of selected massively multilingual collections of monolingual data listed in chronological order. We report size, token counts, language coverage, and the proportion of English content. - indicates that data is not available. * indicates that the English percentage was computed over sentence counts, instead of token counts.

B Parallel and monolingual data statistics

	Raw		Filtered		TMX	
Language	**Sentence Pairs**	**English Words**	**Sentence Pairs**	**English Words**	**Sentence Pairs**	**English Words**
sin_Sinh	929,844	15,647,062	450,122	8,248,007	273,430	5,932,234
npi_Deva	1,058,740	18,514,145	523,022	10,176,931	317,120	7,145,363
xho_Latn	1,223,514	16,524,728	655,790	9,339,359	405,605	5,998,358
mal_Mlym	1,686,113	25,600,791	795,653	12,475,940	547,168	9,656,086
nno_Latn	2,358,129	34,771,540	1,175,108	21,352,540	563,791	10,548,302
mar_Deva	2,067,311	34,952,324	952,116	19,606,305	656,962	15,113,175
guj_Gujr	2,134,977	38,906,708	1,165,483	23,631,881	716,777	16,564,683
kan_Knda	2,354,299	37,451,816	1,238,033	21,344,021	720,157	13,965,655
tel_Telu	2,924,532	46,227,504	1,513,237	25,963,464	902,962	17,487,796
tam_Taml	3,859,610	55,779,718	1,759,372	28,369,233	1,111,471	20,718,487
uzn_Latn	2,791,412	37,715,209	1,571,871	25,823,124	1,159,869	19,667,785
urd_Arab	3,866,815	101,346,427	2,200,602	65,830,839	1,399,893	47,591,409
eus_Latn	5,907,808	79,485,282	2,526,198	38,107,950	1,491,873	24,303,464
epo_Latn	5,664,237	91,081,114	3,190,135	60,141,119	1,521,821	30,986,721
mlt_Latn	7,434,717	114,046,030	2,651,758	50,044,197	1,529,471	32,243,598
kaz_Cyrl	3,827,170	55,027,673	2,628,328	39,138,283	1,943,935	30,216,073
swh_Latn	10,125,330	145,685,653	3,680,151	68,766,541	1,985,899	39,952,916
ben_Beng	6,376,109	106,303,435	3,920,955	70,350,081	2,328,136	49,851,040
isl_Latn	11,929,153	146,981,787	6,624,589	91,089,371	2,694,541	47,440,271
gle_Latn	7,685,880	133,441,028	4,421,130	89,932,030	2,697,582	59,065,530
glg_Latn	8,680,808	145,602,784	5,166,276	99,562,132	2,783,727	58,437,672
bel_Cyrl	11,493,046	154,657,914	6,092,481	90,760,902	3,140,958	50,113,002
azj_Latn	8,506,772	118,079,751	4,765,278	72,026,597	3,188,231	51,425,346
pes_Arab	9,434,306	192,718,387	5,391,049	130,840,005	3,448,296	95,822,037
cym_Latn	9,390,284	156,087,956	6,348,606	125,442,540	3,867,402	82,244,645
afr_Latn	15,901,372	246,703,185	7,452,216	139,249,006	3,987,340	80,857,410
mkd_Cyrl	10,815,504	185,651,668	7,175,217	131,839,062	3,991,617	78,629,993
tha_Thai	13,818,095	102,830,296	7,551,187	52,171,172	4,088,354	34,155,503
als_Latn	11,171,352	208,460,688	6,943,910	145,918,660	4,166,536	94,263,904
bos_Latn	12,480,871	193,998,734	7,527,232	139,457,685	4,559,328	92,723,229
srp_Cyrl	17,605,882	244,921,478	9,618,806	153,171,621	5,291,686	90,518,351
zsm_Latn	47,173,963	558,911,698	22,298,471	301,446,773	8,432,285	147,009,838
heb_Hebr	34,004,891	431,453,938	21,600,460	279,563,405	8,686,089	162,768,846
est_Latn	29,934,421	362,843,780	16,629,846	223,025,816	8,797,574	133,824,400
hin_Deva	26,345,062	500,967,390	16,337,324	372,581,817	9,926,620	263,709,932
slv_Latn	30,956,083	449,919,060	18,290,300	301,699,622	10,336,528	188,709,019
lvs_Latn	39,599,210	476,030,125	24,504,355	316,955,851	11,294,618	183,588,490
lit_Latn	47,035,968	553,786,167	27,879,310	351,354,617	12,881,354	205,285,778
cat_Latn	40,922,098	671,563,410	26,451,844	521,397,424	13,080,859	292,854,267
hrv_Latn	45,617,022	627,929,707	27,783,979	420,953,899	14,263,908	250,103,294
ara_Arab	41,671,896	759,192,353	31,389,602	618,611,620	17,505,366	424,460,900
kor_Hang	76,980,595	865,790,290	46,010,282	530,544,997	18,393,859	294,720,264
jpn_Jpan	105,291,263	575,207,953	48,568,992	178,462,644	18,894,019	80,778,230
vie_Latn	47,831,389	1,077,677,445	35,072,681	831,280,934	19,231,770	502,683,444
slk_Latn	62,840,882	798,323,342	40,704,524	566,060,898	20,056,339	332,818,300
tur_Latn	84,823,944	1,054,338,198	46,493,019	656,237,839	21,616,652	402,325,110
bul_Cyrl	63,059,982	939,666,726	40,936,972	670,141,259	22,725,326	420,768,808
nob_Latn	71,277,875	969,526,777	45,204,041	700,446,147	22,912,722	395,447,564
ukr_Cyrl	68,111,464	862,726,167	46,141,511	637,243,802	25,125,019	400,949,773
fin_Latn	98,138,078	1,028,494,366	59,836,942	667,840,953	29,067,875	383,463,787

Table 5: Statistics for the parallel portion of HPLT v2 before filtering (Raw), after Bicleaner AI (Filtered) and after deduplication (TMX). Languages are in increasing order of deduplicated sentence pairs.

Language	Segments	Tokens	Characters	Documents
ace_Arab	1.170e+02	8.363e+03	4.973e+04	1.600e+01
ace_Latn	2.062e+05	8.196e+06	5.083e+07	1.293e+04
afr_Latn	3.774e+07	1.000e+09	5.947e+09	1.457e+06
als_Latn	9.510e+07	2.713e+09	1.610e+10	5.385e+06
amh_Ethi	7.006e+06	1.959e+08	1.031e+09	2.955e+05
ara_Arab	2.200e+09	4.814e+10	2.795e+11	8.267e+07
asm_Beng	2.677e+06	7.344e+07	4.757e+08	1.757e+05
ast_Latn	7.426e+06	1.950e+08	1.244e+09	2.732e+05
awa_Deva	1.315e+05	6.049e+06	2.877e+07	7.281e+03
ayr_Latn	1.885e+05	3.068e+06	2.508e+07	9.223e+03
azb_Arab	2.389e+06	3.958e+07	2.602e+08	6.611e+04
azj_Latn	1.266e+08	2.569e+09	1.962e+10	6.485e+06
bak_Cyrl	3.139e+06	7.533e+07	5.585e+08	1.708e+05
bam_Latn	9.172e+04	3.982e+06	2.074e+07	5.721e+03
ban_Latn	6.011e+05	1.134e+07	7.724e+07	1.070e+04
bel_Cyrl	4.884e+07	1.212e+09	8.540e+09	2.320e+06
bem_Latn	1.335e+05	4.523e+06	3.232e+07	6.136e+03
ben_Beng	1.760e+08	4.639e+09	3.016e+10	1.104e+07
bho_Deva	4.583e+05	1.347e+07	6.865e+07	2.864e+04
bjn_Arab	1.953e+04	5.482e+05	3.317e+06	1.112e+03
bjn_Latn	3.663e+05	8.048e+06	5.597e+07	1.876e+04
bod_Tibt	4.650e+05	5.781e+06	2.685e+08	2.744e+04
bos_Latn	2.682e+08	7.255e+09	4.607e+10	1.461e+07
bug_Latn	3.855e+04	2.705e+06	1.931e+07	2.023e+03
bul_Cyrl	6.814e+08	1.530e+10	9.693e+10	2.809e+07
cat_Latn	3.833e+08	1.002e+10	6.019e+10	1.855e+07
ceb_Latn	2.865e+06	8.589e+07	5.157e+08	1.388e+05
ces_Latn	1.927e+09	4.208e+10	2.739e+11	7.529e+07
cjk_Latn	3.670e+04	9.647e+05	7.432e+06	1.196e+03
ckb_Arab	5.226e+06	1.426e+08	9.128e+08	2.737e+05
crh_Latn	1.381e+06	3.676e+07	2.811e+08	1.227e+05
cym_Latn	1.557e+07	4.090e+08	2.402e+09	7.581e+05
dan_Latn	8.730e+08	2.120e+10	1.334e+11	3.384e+07
deu_Latn	1.113e+10	2.515e+11	1.782e+12	4.821e+08
dik_Latn	3.465e+04	2.295e+06	1.154e+07	2.325e+03
dyu_Latn	2.456e+04	1.194e+06	5.552e+06	1.390e+03
dzo_Tibt	3.997e+04	4.222e+05	7.375e+06	1.626e+03
ell_Grek	1.849e+09	4.270e+10	2.835e+11	7.033e+07
eng_Latn	1.165e+11	2.862e+12	1.708e+13	4.389e+09
epo_Latn	2.035e+07	4.716e+08	2.976e+09	8.189e+05
est_Latn	2.644e+08	4.742e+09	3.602e+10	8.449e+06
eus_Latn	3.762e+07	7.767e+08	6.052e+09	1.974e+06
ewe_Latn	1.434e+05	4.308e+06	2.132e+07	3.772e+03
fao_Latn	4.526e+06	9.345e+07	5.818e+08	2.399e+05
fij_Latn	1.789e+05	7.263e+06	3.769e+07	8.914e+03
fin_Latn	9.766e+08	1.845e+10	1.557e+11	3.482e+07
fon_Latn	1.476e+04	1.233e+06	5.335e+06	1.226e+03
fra_Latn	1.056e+10	2.370e+11	1.457e+12	4.018e+08
fur_Latn	7.300e+05	2.082e+07	1.147e+08	3.667e+04
fuv_Latn	1.340e+05	5.143e+06	2.990e+07	7.760e+03
gaz_Latn	9.736e+05	2.888e+07	2.192e+08	4.914e+04
gla_Latn	3.307e+06	8.066e+07	4.836e+08	1.374e+05
gle_Latn	1.099e+07	2.957e+08	1.749e+09	4.908e+05
glg_Latn	6.118e+07	1.639e+09	1.011e+10	3.020e+06
grn_Latn	1.713e+06	3.072e+07	2.186e+08	7.342e+04
guj_Gujr	2.064e+07	5.768e+08	3.386e+09	1.134e+06

Table 6: Counts of segments, tokens, characters and documents for each language in the monolingual HPLT v2 datasets. Tokens are words as defined by Unix `wc`.

Language	Segments	Tokens	Characters	Documents
hat_Latn	4.635e+06	1.223e+08	6.389e+08	2.127e+05
hau_Latn	5.688e+06	1.526e+08	8.535e+08	3.159e+05
heb_Hebr	4.666e+08	9.966e+09	5.682e+10	1.712e+07
hin_Deva	2.674e+08	8.637e+09	4.396e+10	1.365e+07
hne_Deva	5.500e+04	2.199e+06	1.059e+07	2.806e+03
hrv_Latn	2.971e+08	7.307e+09	4.800e+10	1.230e+07
hun_Latn	1.419e+09	3.052e+10	2.252e+11	5.187e+07
hye_Armn	6.524e+07	1.405e+09	1.072e+10	3.599e+06
ibo_Latn	1.411e+06	3.829e+07	2.052e+08	5.629e+04
ilo_Latn	1.120e+06	2.478e+07	1.568e+08	4.875e+04
ind_Latn	2.389e+09	5.462e+10	3.842e+11	9.814e+07
isl_Latn	6.964e+07	1.536e+09	9.593e+09	2.841e+06
ita_Latn	5.127e+09	1.274e+11	8.206e+11	2.218e+08
jav_Latn	6.431e+06	1.378e+08	9.375e+08	1.960e+05
jpn_Jpan	2.327e+10	4.236e+10	9.011e+11	4.177e+08
kab_Latn	3.452e+05	9.222e+06	5.419e+07	1.510e+04
kac_Latn	1.594e+05	5.955e+06	2.840e+07	7.587e+03
kam_Latn	1.426e+04	6.740e+05	4.645e+06	1.183e+03
kan_Knda	2.493e+07	5.329e+08	4.298e+09	1.336e+06
kas_Arab	2.711e+04	6.780e+05	3.468e+06	9.490e+02
kas_Deva	1.357e+03	3.194e+04	1.854e+05	1.060e+02
kat_Geor	6.372e+07	1.244e+09	1.016e+10	3.335e+06
kaz_Cyrl	8.101e+07	1.409e+09	1.113e+10	2.637e+06
kbp_Latn	4.679e+04	4.258e+06	2.090e+07	7.075e+03
kea_Latn	4.391e+04	1.143e+06	6.144e+06	1.962e+03
khk_Cyrl	5.347e+07	1.342e+09	9.327e+09	2.121e+06
khm_Khmr	9.864e+06	1.138e+08	2.122e+09	7.010e+05
kik_Latn	5.193e+04	1.428e+06	9.292e+06	3.995e+03
kin_Latn	1.917e+06	5.074e+07	3.671e+08	9.270e+04
kir_Cyrl	1.004e+07	2.467e+08	1.925e+09	6.761e+05
kmb_Latn	1.180e+04	3.831e+05	2.068e+06	5.310e+02
kmr_Latn	7.147e+06	1.959e+08	1.123e+09	3.643e+05
knc_Arab	1.083e+04	2.620e+05	1.302e+06	2.450e+02
knc_Latn	1.052e+04	2.409e+06	1.195e+07	2.472e+03
kon_Latn	4.748e+04	1.944e+06	1.127e+07	2.542e+03
kor_Hang	1.358e+09	1.970e+10	8.923e+10	3.887e+07
lao_Laoo	3.200e+05	5.178e+06	8.468e+07	2.950e+04
lij_Latn	1.577e+05	5.593e+06	3.146e+07	8.371e+03
lim_Latn	7.140e+06	1.806e+08	1.125e+09	3.679e+05
lin_Latn	2.003e+05	5.555e+06	3.292e+07	7.588e+03
lit_Latn	3.222e+08	6.676e+09	5.039e+10	1.334e+07
lmo_Latn	2.125e+06	5.964e+07	3.454e+08	1.462e+05
ltg_Latn	1.514e+05	3.790e+06	2.688e+07	9.209e+03
ltz_Latn	5.059e+06	1.072e+08	7.104e+08	2.469e+05
lua_Latn	3.869e+04	1.368e+06	9.005e+06	1.083e+03
lug_Latn	4.075e+05	9.176e+06	6.796e+07	2.128e+04
luo_Latn	8.412e+04	3.727e+06	2.033e+07	4.153e+03
lus_Latn	3.433e+06	1.252e+08	6.520e+08	1.604e+05
lvs_Latn	1.738e+08	3.461e+09	2.518e+10	6.772e+06
mag_Deva	1.929e+04	8.906e+05	4.283e+06	3.280e+02
mai_Deva	6.455e+05	1.779e+07	9.674e+07	2.498e+04
mal_Mlym	4.800e+07	9.737e+08	9.489e+09	3.105e+06
mar_Deva	3.632e+07	9.807e+08	6.622e+09	2.080e+06
min_Latn	6.008e+05	1.098e+07	7.477e+07	2.504e+04
mkd_Cyrl	5.701e+07	1.485e+09	9.440e+09	3.566e+06
mlt_Latn	8.675e+06	1.958e+08	1.442e+09	3.673e+05

Table 6: Counts of segments, tokens, characters and documents for each language in the monolingual HPLT v2 datasets. Tokens are words as defined by Unix `wc`.

Language	Segments	Tokens	Characters	Documents
mni_Beng	6.576e+04	1.627e+06	1.179e+07	2.934e+03
mos_Latn	1.910e+04	8.075e+05	3.864e+06	9.310e+02
mri_Latn	2.795e+06	8.676e+07	4.243e+08	1.083e+05
mya_Mymr	3.050e+07	4.532e+08	5.819e+09	1.368e+06
nld_Latn	3.075e+09	7.141e+10	4.511e+11	1.387e+08
nno_Latn	3.460e+07	8.603e+08	5.404e+09	1.423e+06
nob_Latn	6.760e+08	2.154e+10	1.332e+11	2.705e+07
npi_Deva	3.714e+07	1.128e+09	7.256e+09	2.778e+06
nso_Latn	1.433e+05	5.322e+06	2.749e+07	6.066e+03
nus_Latn	8.514e+03	3.932e+05	1.882e+06	2.720e+02
nya_Latn	1.344e+06	2.706e+07	2.029e+08	5.312e+04
oci_Latn	4.195e+06	1.027e+08	6.354e+08	1.899e+05
ory_Orya	3.596e+06	1.201e+08	7.815e+08	4.129e+05
pag_Latn	8.583e+04	5.657e+06	3.352e+07	6.900e+03
pan_Guru	1.174e+07	3.722e+08	1.902e+09	5.846e+05
pap_Latn	1.387e+06	4.671e+07	2.541e+08	8.981e+04
pbt_Arab	8.455e+06	2.794e+08	1.304e+09	4.665e+05
pes_Arab	3.963e+09	8.855e+10	4.551e+11	9.050e+07
plt_Latn	4.736e+06	1.171e+08	8.103e+08	2.078e+05
pol_Latn	4.461e+09	8.953e+10	6.316e+11	1.754e+08
por_Latn	6.125e+09	1.463e+11	8.965e+11	2.378e+08
prs_Arab	6.900e+07	1.844e+09	9.567e+09	2.839e+06
quy_Latn	4.943e+05	1.731e+07	1.434e+08	3.694e+04
ron_Latn	1.697e+09	4.005e+10	2.507e+11	6.588e+07
run_Latn	1.752e+06	4.444e+07	3.165e+08	1.373e+05
rus_Cyrl	2.629e+10	5.409e+11	3.908e+12	8.847e+08
sag_Latn	5.190e+04	3.612e+06	1.674e+07	3.161e+03
san_Deva	3.281e+06	4.380e+07	3.592e+08	5.491e+04
sat_Olck	4.580e+04	1.085e+06	6.266e+06	2.566e+03
scn_Latn	1.650e+06	4.239e+07	2.523e+08	8.197e+04
shn_Mymr	9.214e+04	1.648e+06	2.121e+07	6.003e+03
sin_Sinh	3.371e+07	7.956e+08	4.981e+09	1.153e+06
slk_Latn	4.943e+08	1.063e+10	7.037e+10	2.183e+07
slv_Latn	2.386e+08	5.435e+09	3.526e+10	1.028e+07
smo_Latn	1.012e+06	3.709e+07	1.861e+08	4.586e+04
sna_Latn	1.202e+06	2.392e+07	1.926e+08	6.108e+04
snd_Arab	2.826e+06	8.953e+07	4.286e+08	1.003e+05
som_Latn	1.638e+07	3.888e+08	2.565e+09	9.665e+05
sot_Latn	1.085e+06	3.100e+07	1.715e+08	4.392e+04
spa_Latn	1.212e+10	3.220e+11	1.954e+12	5.031e+08
srd_Latn	9.171e+05	2.389e+07	1.487e+08	5.382e+04
srp_Cyrl	9.381e+07	2.519e+09	1.616e+10	4.123e+06
ssw_Latn	6.213e+04	9.943e+05	8.821e+06	2.036e+03
sun_Latn	3.238e+06	6.963e+07	4.753e+08	1.148e+05
swe_Latn	1.755e+09	4.011e+10	2.511e+11	6.681e+07
swh_Latn	3.431e+07	7.177e+08	4.664e+09	1.374e+06
szl_Latn	6.366e+05	1.468e+07	1.038e+08	4.093e+04
tam_Taml	1.686e+08	2.981e+09	2.624e+10	6.106e+06
taq_Latn	1.388e+04	1.544e+06	8.845e+06	1.747e+03
tat_Cyrl	1.345e+07	2.967e+08	2.157e+09	6.307e+05
tel_Telu	3.919e+07	8.354e+08	6.505e+09	2.058e+06
tgk_Cyrl	2.485e+07	6.248e+08	4.590e+09	1.261e+06
tgl_Latn	5.288e+07	1.346e+09	8.131e+09	1.869e+06
tha_Thai	3.391e+08	3.506e+09	5.998e+10	1.770e+07
tir_Ethi	1.128e+06	3.672e+07	1.816e+08	6.469e+04
tpi_Latn	2.824e+05	1.251e+07	6.453e+07	1.398e+04

Table 6: Counts of segments, tokens, characters and documents for each language in the monolingual HPLT v2 datasets. Tokens are words as defined by Unix `wc`.

Language	Segments	Tokens	Characters	Documents
tsn_Latn	1.322e+05	5.273e+06	2.767e+07	6.050e+03
tso_Latn	2.212e+05	8.668e+06	4.929e+07	1.101e+04
tuk_Latn	3.355e+06	7.068e+07	5.700e+08	1.710e+05
tum_Latn	9.901e+04	2.876e+06	2.110e+07	4.384e+03
tur_Latn	2.575e+09	5.167e+10	3.896e+11	1.166e+08
twi_Latn	1.256e+05	4.696e+06	2.418e+07	5.860e+03
uig_Arab	8.982e+06	2.239e+08	1.747e+09	4.424e+05
ukr_Cyrl	1.169e+09	2.523e+10	1.829e+11	4.740e+07
umb_Latn	5.991e+04	2.431e+06	1.541e+07	2.471e+03
urd_Arab	5.063e+07	2.126e+09	1.001e+10	3.194e+06
uzn_Latn	1.480e+07	3.513e+08	2.846e+09	7.069e+05
vec_Latn	1.579e+06	3.526e+07	2.180e+08	8.480e+04
vie_Latn	3.020e+09	8.320e+10	3.795e+11	1.007e+08
war_Latn	2.009e+05	5.889e+06	3.557e+07	1.387e+04
wol_Latn	1.615e+05	5.463e+06	2.754e+07	5.679e+03
xho_Latn	1.821e+06	3.034e+07	2.587e+08	6.309e+04
ydd_Hebr	2.940e+06	7.753e+07	4.585e+08	1.283e+05
yor_Latn	1.469e+06	4.281e+07	2.178e+08	6.613e+04
yue_Hant	1.235e+06	3.268e+06	7.430e+07	6.129e+04
zho_Hans	4.245e+10	7.403e+10	2.352e+12	1.247e+09
zho_Hant	4.480e+09	9.510e+09	2.868e+11	1.571e+08
zsm_Latn	5.798e+08	1.148e+10	7.843e+10	1.842e+07
zul_Latn	2.710e+06	4.436e+07	3.808e+08	1.136e+05

Table 6: Counts of segments, tokens, characters and documents for each language in the monolingual HPLT v2 datasets. Tokens are words as defined by Unix `wc`.

C Sources of web crawls

Name	Years	Size (TB)
IA full crawls	**2012–2020**	**3390**
wide5	2012–2012	365
wide6	2012–2013	204
wide10	2014–2014	91
wide11	2014–2014	420
wide12	2015–2015	449
wide15	2016–2017	358
wide16	2017–2018	768
wide17	2018–2020	641
survey3	2015–2016	94
CC full crawls	**2014–2022**	**743**
CC-MAIN-2014-35	2014	43
CC-MAIN-2014-42	2014	54
CC-MAIN-2015-11	2015	29
CC-MAIN-2015-48	2015	30
CC-MAIN-2017-04	2017	54
CC-MAIN-2018-05	2018	75
CC-MAIN-2018-22	2018	52
CC-MAIN-2018-43	2018	59
CC-MAIN-2021-43	2021	86
CC-MAIN-2022-27	2022	85
CC-MAIN-2022-40	2022	83
CC-MAIN-2022-49	2022	93
Partial crawls	**2013-2023**	**317**
1% of WARCs from 81 CC crawls	2013-2023	46
7% of IA ArchiveBot	2013-2023	271

Table 7: List of web crawls used to construct HPLT v2. From IA, we use 8 Wide crawls, 1 Survey crawl containing main pages of websites and a random sample of 7% of items from IA ArchiveBot. From CC, we use 12 randomly-selected full crawls, plus a 1% sample of WARCs from each of the other 81 available crawls.

D Yields of different crawls

To figure out how different web crawls contribute to our datasets and which crawls are the most promising sources of monolingual corpora in general, we compared crawls from two points of view: the amount of texts extracted from each crawl and the quality of these texts. In this section, we study crawls from the first point of view, while in H the results of manual quality inspection are presented.

To make a comparison, we group all crawls into groups according to their age and source. The oldest IA wide crawls from 2012-2014 (from wide5 up to wide11) are assigned to the group `ia_o`, the newest wide16, wide17 crawls from 2017-2020 to the group `ia_n`, and the wide12, wide15 crawls in the middle to the group `ia_m`. CC crawls are split by age following the same time periods, but additionally a group `cc_r` is introduced for the recent CC crawls from 2021-2023 (we don't have IA wide crawls from this time period). Finally, the IA survey3 and ArchiveBot crawls form their own groups `ia_survey` and `ia_archivebot`. In total, we have 9 groups of crawls.

For different processing stages, Figure 6 visualizes how much data comes from different groups of crawls. While originally less than 20% of our crawls are CC crawls, they contribute about half of the raw text before duplication and more than 60% of the text after deduplication and cleaning. Especially high-yielding are the new and recent CC crawls, they are only 6% and 8% of all crawls in size but contribute 28% and 30% of text (both when counting in characters and in documents) to the cleaned version. On the other hand, the newest IA wide crawls are 32% of all crawls in size but contribute only about 11% of text.

Figure 7 suggests another point of view showing yields for different crawls, or more specifically, how much text (measured in the number of characters) is extracted from 1 GB of compressed WARC files for each crawl. Evidently, CC crawls have the highest yields, especially the newer ones. Compared to the newer CC crawls, for the older CC crawls more data is filtered during deduplication and cleaning, giving finally lower yields despite a bit higher yields of raw texts. IA wide crawls have 4-8x smaller yields than CC crawls. The survey IA crawl has a comparable yield to the wide crawls in the final dataset. Since they are publicly available, it probably makes sense to employ more of these crawls in the future. Finally, the ArchiveBot IA crawl has remarkably low yields.

Despite having a lower contribution in general, for some languages, IA crawls supply the majority of texts. Figure 8 shows 15 languages with the highest proportion of texts from IA crawls. They include both high-resourced (Chinese, Western Persian) and low-resourced languages. Deduplication and cleaning significantly reduce the number of languages with high contribution of IA. For instance, before deduplication and cleaning there are 49 languages having more than 70% of texts (characters) coming from IA and only 6 such languages after.

E Frequent *n*-grams

We obtain frequent n-grams (up to order 5) in each dataset after tokenizing text and applying some restrictions:

- n-grams must start and end in the same seg-

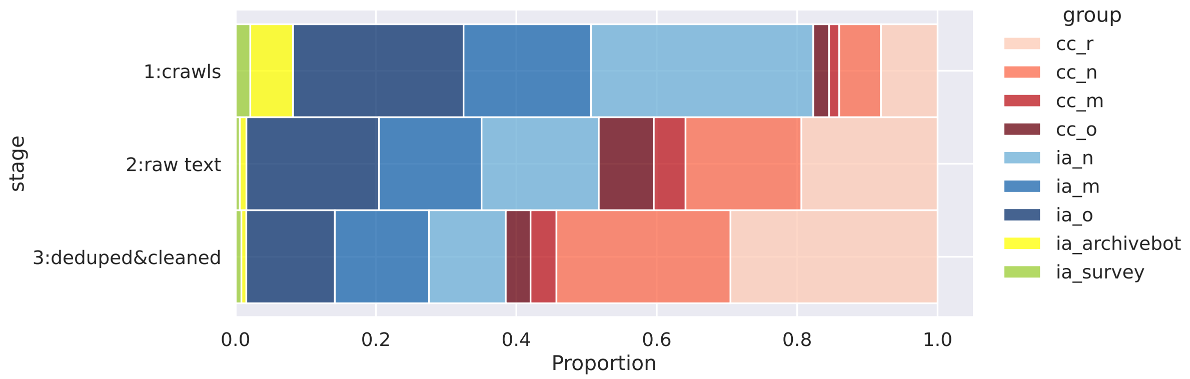

Figure 6: Proportions of data from different groups of crawls at various processing stages. Crawls were quantified in TB of compressed WARC files, while raw texts and deduplicated cleaned texts in characters.

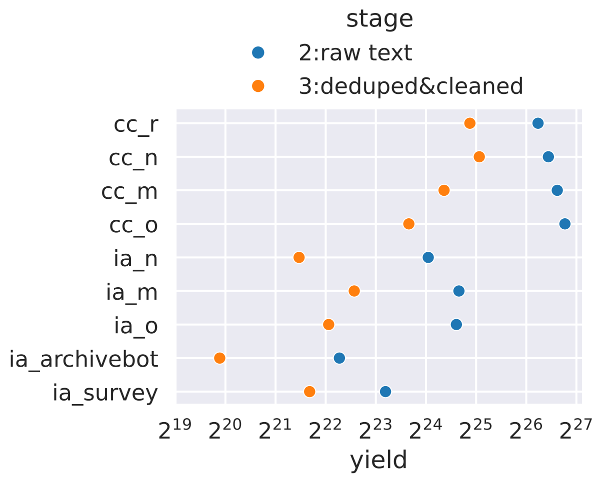

Figure 7: Yields (in characters per 1 GB of raw compressed crawls) of different crawls at different stages.

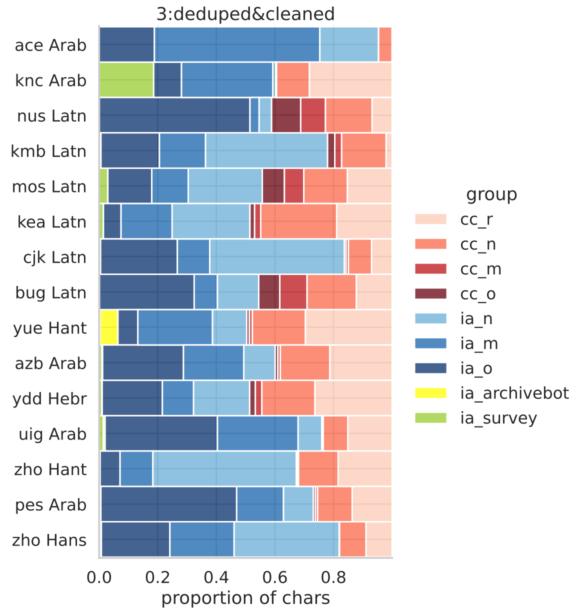

Figure 8: Proportions of texts from different groups of crawls for the 15 languages with the largest contribution of IA crawls.

ment (i.e. no line breaks are allowed in the middle of a n-gram)

- n-grams containing any punctuation are discarded
- n-grams that start or end in stopwords are discarded
- n-grams are calculated case-insensitive
- all tokens in the n-gram must have at least one alphabetic character

Tables 8 and 9 show the 5 most frequent n-grams (orders 1 to 5) in HPLT v2. In the case of parallel datasets, n-grams are selected from the target (non-English) side of the segments. Translation to English is obtained with Google Translate.[30]

We find that most datasets (both monolingual and parallel) contain frequent n-grams that seem to be boilerplate, such as "edit source", "read more", "click button" or "view map". This kind of content usually comes from Wikipedia and Blogspot. In the monolingual datasets, there is a large amount of text that seems to come from headers or footers in news webpages, e.g. "latest news". Biblical n-grams (such as "god" or "jehovah") are also very frequent in some datasets, notably African languages, matching our observations about frequent domains (Appendix F). Some frequent n-grams suggest poor-quality content in some datasets, since they seem to be related to downloads webpages, online game platforms or betting sites.

For the parallel datasets, we observe that on the English side the frequent n-grams are very similar across all languages. For the languages with the most data, hotels and legal notices are the most

[30]https://translate.google.com/

common kind of n-grams. The smaller parallel datasets tend to exhibit more variety of n-grams and include n-grams alluding to political leaders or city names, which suggest more locally-generated content (probably from news sites). Finally, frequent n-grams in parallel datasets from Eastern European countries usually contain mentions to European institutions (such as the European Parliament or the European Commission). This matches our observations on TLDs in Appendix G.

F Frequent domains

Inspecting the most common domain names in the datasets is one way to understand the type of content we can find in it. Table 10 gives the datasets with the highest proportion of frequent domain classes, and Table 11 gives the datasets with the highest proportion of frequent domain classes for the parallel data. We make the following general observations:

- Mid-to-large-sized datasets show a wider variety of domains with no clear majority source. However, in monolingual datasets, blogging platforms usually get a significant portion of the total (Table 10).
- Wikipedia tends to be among the most frequent domains for both monolingual and parallel datasets. It is usually the most frequent domain for smaller language datasets (Table 10).
- Hotel and travel webpages are much more frequent in the larger parallel datasets and very infrequent in the monolingual data (Table 11).
- News and media outlets are also a frequent content source in monolingual datasets, with some news websites getting a significant percentage in different datasets: for example, regional websites from the Free Radio Europe[31] or Voice of America[32] networks (Table 10).
- Religious and biblical content is also very frequent in the smaller monolingual and parallel datasets. This is specially notable in the case of African languages, which often get more than three quarters of their content from such sites (Table 10).
- Software and online gaming websites are usually among the top-10 most frequent domains in almost all parallel datasets.
- Chinese shopping websites are common in the larger parallel datasets of non-European languages (Vietnamese, Japanese, Korean, Arabic and Turkish).
- No pornographic webpages appear in the top domains, implying our filter for such content worked as expected.

G Geographic TLDs

Tables 12 and 13 list the most frequent examples of geographic TLDs in the monolingual and parallel HPLT v2 corpora respectively. We make the following observations in addition to those made in the main text:

- In general, the most frequent TLDs in many of the datasets are generic (such as `.com`, `.org` or `.info`).
- Some TLDs are frequent because they "sound good" rather than indicating the kind of content or language: `.icu` (because it reads like "I see you"), `.is` (official TLD for Iceland, but used as a verb and very noticeably in `bible.is`, a religious webpage whose domain is usually in the top-10 most frequent TLDs), `.tv` (for the country of Tuvalu, but widely used for TV-related web pages), `.co` (for Colombia, but mostly used for companies), `.no` (for Norway, but used as a negative particle), `.nu` (for the island of Niue, but used because it sounds like "new"), etc.
- There are common TLDs for super-national territories: `.eu` (European Union), `.africa`, `.asia`, etc.
- In our monolingual datasets, there is frequently one geographic TLD among the 10 most frequent ones that clearly surpasses the others. The "winning" TLD is usually from the country where the dataset language is spoken most, indicating that the text content is probably in the correct language. The percentage of this "winning country" varies depending on the amount of general purpose TLD in the dataset, but it is in general higher for European countries. This "winning" geographic TLD, in the case of parallel datasets, is less frequent and, when present, its portion of the total is noticeably lower than for the monolingual datasets.
- Many African languages do not have a significant portion of geographic TLDs (beyond the aforementioned `.is`, `.no`, etc).
- For some languages, there are a few coun-

[31] https://www.rferl.org/navigation/allsites
[32] https://www.voanews.com/navigation/allsites

n-gram (original)	n-gram (translated)	Dataset	Occurrences
Blogging & social networks boilerplate			
read more	-	Tosk Albanian	318,636
read more	-	Malayalam	244,935
read more	-	Telugu	225,178
posted by	-	Malayalam	177,507
read more	-	Nepali	176,500
മലയാളത്തിലോ ഇംഗ്ലീഷിലോ	readers' comments	Malayalam	414,905
you must be logged in	-	Tagalog	138,209
лістапад кастрычнік верасень	november october september	Belarusian	126,015
Wikipedia boilerplate			
editar	edit	Galician	3,177,972
redakti	edit	Esperanto	3,013,130
editar a fonte	edit source	Galician	1,491,606
redakti fonton	edit source	Esperanto	1,405,770
aldatu iturburu kodea	edit source code	Basque	921,847
endre wikiteksten	edit wiki text	Norwegian Nynorsk	777,797
правіць зыходнік	edit source	Belarusian	710,669
modificar la font	edit source	Occitan	568,442
editar la fonte	edit source	Asturian	567,301
խմբագրել կոդը	edit source text	Armenian	517,252
წყაროს რედაქტირება	edit source	Georgian	355,269
уреди извор	edit source	Macedonian	506,624
wysig bron	edit source	Afrikaans	345,356
quelltext änneren	edit source	Luxembourgish	278,814
Religious & biblical content			
uyehova	jehovah	Xhosa	47,227
yehova	jehovah	Tumbuka	27,729
chiuta	god	Tumbuka	23,684
yehowa	jehovah	Ewe	16,739
biblia	bible	Ewe	10,777
yehova	jehovah	Chokwe	9,174
yesu	jesus	Chokwe	6,984
nin diyos	by god	Pangasinan	6,636
yeova	jehovah	Kamba	3,785
Low-quality content indicators			
ਪੋਰਨ ਵੀਡੀਓ	porn video	Punjabi	143,816
piala donya	world cup	Javanese	121,060
piala dunya	world cup	Sundanese	93,086
compartir descargar reproducir	share download play	Asturian	71,042
ndajnë këtë lojë me miqtë	share this game with friends	Tosk Albanian	67,888
tohan maén bal	soccer betting	Javanese	52,873
luaj online flash lojë	play online flash game	Tosk Albanian	47,112
fifa world cup	-	Khmer	45,367
qatar world cup	-	Khmer	35,895
gêm hon gyda 'ch ffrindiau	(share) this game with your friends	Welsh	12,984
jwèt sou entènèt	online game	Haitian Creole	12,686
play ar líne a flash	play flash online	Irish	10,954
सेक्सी मूवी	sexy movie	Bhojpuri	8,812
tohan piala dunya	world cup betting	Sundanese	6,691

Table 8: Frequent n-grams in monolingual datasets.

n-gram (original)	n-gram (translated)	Dataset	Occurrences
Legal boilerplate			
osobných údajov	personal data	Slovak	776.876
лични данни	personal data	Bulgarian	487.125
osobnih podataka	personal data	Croatian	290.381
osebnih podatkov	presonal data	Slovene	244.755
asmens duomen	personal data	Lithuanian	239.844
personas datu	personal data	Latvian	202.714
персональних даних	personal data	Ukrainian	153.272
ข้อมูล ส่วนบุคคล	personal information	Thai	134.876
personas datus	personal data	Latvian	128.843
Hotels & travels			
hotel	hotel	Malay	681.806
המלון	hotel	Hebrew	633.416
bilik	rooms	Malay	295.982
ค้นหา โรงแรม	find hotels	Thai	152.206
tripadvisor er stolt	tripadvisor is proud	Norwegian Bokmål	145.267
จอง ห้องพัก	book rooms	Thai	136.329
tatil hem de i seyahatleri	holidays and business travel	Turkish	122.057
hótel	hotels	Icelandic	113.560
kävelymatkan päässä	walking distance	Finnish	103.208
European Union			
euroopa liidu	european union	Estonian	45.900
eiropas parlamenta un padomes	european parliament and council	Latvian	33.524
европейския парламент и на съвета	european parliament and council	Bulgarian	31.860
europos parlamento ir tarybos	european parliament and council	Lithuanian	30.219
evropskega parlamenta in sveta	of the european parliament and council	Slovene	27.164
euroopan parlamentin ja neuvoston	european parliament and council	Finnish	24.667
tal-parlament ewropew	of the european parliament	Maltese	21.479
Boilerplate			
הצג מפה	show map	Hebrew	344.614
지도 보 기	show map	Korean	304.644
haritay göster	show map	Turkish	191.672
اعرض الخريطة	show map	Arabic	187.615
kullancsndan yeniden bloglad	reblogged from	Turkish	180.189
עריכת קוד מקור	edit code	Hebrew	83.767
modifica el codi	edit code	Catalan	64.734
editar a fonte	edit source	Galician	54.873
редактиране на кода	edit code	Bulgarian	46.778
लिंक पर क्लिक	click link	Hindi	30.258
Religious & biblical			
mungu	god	Swahili	166.221
uyehova	jehovah	Xhosa	24.281
heilige gees	holy spirit	Afrikaans	14.284
thotë zoti	says god	Albanian	13.480
uyesu	jesus	Xhosa	9.135
diras la eternulo	says the lord	Esperanto	7.425
Software & games			
permainan dalam talian	online game	Malay	19.342
این نرم افزار	this software	Farsi	18.992
call of duty	call of duty	Farsi	18.765
mchezo huu	this game	Swahili	16.014
gêm ar-lein	online game	Welsh	11.774
luaj online flash	play online flash	Albanian	9.298
споделување на играта	sharing the game	Macedonian	8.170

Table 9: Frequent n-grams in parallel datasets (non-English side).

Dataset	% of documents
Blogging platforms	
Standard Malay	50%
Magahi	48%
Greek	24%
Cantonese	22%
Portuguese	16%
Finnish	16%
Swedish	13%
Spanish	10%
Wikipedia	
Santali	90%
Ligurian	80%
Waray	74%
Iloko	66%
Esperanto	66%
Occitan	62%
Sicilian	55%
News & media	
Crimean Tatar	61%
Tigrinya	48%
Banjar (Arabic)	46%
Nigerian Fulfulde	38%
Turkmen	32%
Kyrgyz	30%
Rundi	29%
Religious & biblical	
Dyula	99%
Fon	96%
Bemba	95%
Tumbuka	94%
Kamba	93%
Chokwe	92%
Central Kanuri (Latin)	91%
Luba-Lulua	88%
Sango	88%
Umbundu	84%

Table 10: Languages with the biggest proportion of frequent domain classes in the monolingual HPLT v2 corpora.

tries TLDs in the top-10 from closely related countries or territories (for example, from former colonial rulers (i.e. African languages datasets) or with geostrategic interests (i.e. `.ru` (Russia) appearing in all former Soviet states). This may indicate "language contamination" in the data.

H Manual quality inspection

In this section, we study how the quality of the extracted texts varies between older and newer crawls, and also between IA and CC crawls. More specifically, for a particular language we wanted to understand if there are any substantial differences in the proportions of texts classified as this language by mistake or just undesirable texts.

Language	% of segments
Hotels & travels	
Icelandic	42%
Malay	34%
Hebrew	26%
Lithuanian	21%
Korean	18%
Thai	16%
Norwegian Bokmål	16%
Japanese	15%
Wikipedia	
Norwegian Nynorsk	72%
Galician	37%
Esperanto	36%
Kannada	22%
Macedonian	21%
Telugu	19%
Catalan	19%
Religious & biblical	
Xhosa	70%
Esperanto	28%
Swahili	20%
Nepali	16%
Icelandic	14%
Albanian	14%

Table 11: Frequent domain classes in parallel HPLT v2 datasets for different languages (non-English side).

For this study, we carried out manual annotation of documents from the cleaned version of our dataset asking our annotators to provide three binary annotations for each document.

- **LID ok:** 0 if most of the text is not in the target language, otherwise 1;
- **Unnatural:** 1 if most of the text looks unnatural (e.g. word lists for SEO, mostly boilerplate, etc.), otherwise leave empty;
- **Porn:** 1 if the text looks like pornographic content, otherwise leave empty.

We compared four groups of crawls: among wide IA crawls and CC crawls separately we selected old crawls from 2012-2014 and new crawls from 2017-2020. Among languages spoken by the paper authors, 22 languages were selected for annotation.

For each language and each group of crawls, 50 random documents from the cleaned version of our datasets were annotated by a native or a fluent speaker of this language. In total, 200 documents for each language were annotated, except for Russian where three native speakers annotated 600 documents. Only texts extracted from the documents were shown to the annotators, they did not know which crawl each text came from or any other

Language	% of documents	TLD	Country or Territory
One geographic TLD			
Manipuri	80%	.in	India
Lithuanian	79%	.lt	Lithuania
Polish	77%	.pl	Poland
Hungarian	76%	.hu	Hungary
Danish	76%	.dk	Denmark
Icelandic	73%	.is	Iceland
Faroese	73%	.fo	Faroe Islands
Macedonian	73%	.mk	N. Macedonia
Latgalian	73%	.lv	Latvia
Latvian	72%	.lv	Latvia
Related territories			
Slovak	77%	.sk	Slovakia
	3%	.cz	Czechia
Kazakh	71%	.kz	Kazakhstan
	3%	.ru	Russia
Russian	65%	.ru	Russia
	5%	.ua	Ukraine
	2%	.by	Belarus
Croatian	47%	.hr	Croatia
	3%	.ba	Bosnia
	3%	.rs	Serbia
Kyrgyz	33%	.kg	Kyrgyzstan
	7%	.ru	Russia
Bosnian	29%	.rs	Serbia
	13%	.ba	Bosnia
Language variants			
Romanian	72%	.ro	Romania
	3%	.md	Moldova
Dutch	66%	.nl	Netherlands
	12%	.be	Belgium
German	60%	.de	Germany
	6%	.at	Austria
	5%	.ch	Switzerland
Portuguese	45%	.br	Brazil
	9%	.pt	Portugal
Lombard	47%	.ch	Switzerland
	5%	.it	Italy
Uyghur	36%	.cn	China
	3%	.kz	Kazakhstan
French	30%	.fr	France
	3%	.be	Belgium
	2%	.ca	Canada
	2%	.ch	Switzerland
Spanish	15%	.es	Spain
	4%	.ar	Argentina
	4%	.mx	Mexico
	2%	.cl	Chile
	1%	.pe	Peru

Table 12: Frequent geographic TLDs in monolingual HPLT v2 datasets for different languages.

Language	% of segments	TLD	Country or Territory
One geographic TLD			
Norwegian Nynorsk	36%	.no	Norway
Norwegian Bokmål	34%	.no	Norway
Azerbaijani	33%	.az	Azerbaijan
Macedonian	25%	.mk	North Macedonia
Vietnamese	23%	.vn	Vietnam
Farsi	22%	.ir	Iran
Hebrew	20%	.il	Israel
Sinhala	19%	.lk	Sri Lanka
Serbian	16%	.rs	Serbia
Malay	15%	.my	Malaysia
Hindi	15%	.in	India
Japanese	15%	.jp	Japan
Korean	15%	.kr	South Korea
Related territories (European Union)			
Maltese	68%	.eu	European Union
	4%	.mt	Malta
Slovene	31%	.si	Slovenia
	17%	.eu	European Union
Estonian	35%	.ee	Estonia
	16%	.eu	European Union
Latvian	30%	.lv	Latvia
	16%	.eu	European Union
Lithuanian	35%	.lt	Lithuania
	14%	.eu	European Union
Slovak	44%	.sk	Slovakia
	11%	.eu	European Union
	4%	.cz	Czechia
Croatian	26%	.hr	Croatia
	10%	.eu	European Union
	2%	.ba	Bosnia
Bulgarian	26%	.bg	Bulgaria
	10%	.eu	European Union
Irish	20%	.ie	Ireland
	10%	.eu	European Union
Finnish	44%	.fi	Finland
	7%	.eu	European Union

Table 13: Frequent geographic TLDs in the parallel HPLT v2 datasets for different languages (non-English side).

Language	% Porn ↓	% Unnat. ↓	% LID ↑
Arabic	0 (-)	9 (5-13)	100 (-)
Asturian	0 (-)	28 (22-35)	69 (62-75)
Bengali	1 (-)	0 (-)	100 (-)
Catalan	0 (-)	14 (9-19)	99 (-)
Czech	0 (-)	9 (4-13)	100 (-)
Chinese	0 (-)	25 (18-31)	99 (-)
Dutch	1 (-)	5 (-)	100 (-)
English	1 (-)	13 (8-18)	100 (-)
Finnish	1 (-)	4 (-)	100 (-)
German	1 (-)	2 (-)	98 (-)
Hindi	2 (-)	2 (-)	98 (-)
Iran. Persian	0 (-)	25 (18-31)	99 (-)
Marathi	0 (-)	6 (-)	97 (-)
Modern Greek	0 (-)	3 (-)	100 (-)
Nor. Bokmål	2 (-)	8 (4-11)	99 (-)
Nor. Nynorsk	0 (-)	3 (-)	93 (-)
Polish	1 (-)	7 (3-11)	100 (-)
Russian	2 (1-3)	18 (15-21)	98 (-)
Scot. Gaelic	0 (-)	3 (-)	89 (85-93)
Slovak	0 (-)	10 (6-14)	100 (-)
Spanish	1 (-)	9 (5-13)	100 (-)
Turkish	6 (-)	10 (5-14)	99 (-)

Table 14: Manual quality inspection of a random sample of documents from the cleaned version, stratified by crawls groups. Percentages of extracted texts considered as pornography (% Porn), unnatural texts (% Unnat.), and texts correctly classified by language identification (% LID) (the 95% confidence intervals for the percentage estimates are given in brackets when applicable).

meta-information. For documents longer than 1000 characters, the first 500 characters and 500 characters from the beginning of the second half were shown.

Table 14 shows the results for the four groups combined together.[33] We see that the proportion of pornographic content is low, usually between 0-2% with a maximum of 6% for Turkish. The precision of our LID model for the inspected languages is above 97%, with a few notable exceptions. The worst precision is for Asturian where we observed about 30% of texts being in Spanish or other Spanish minority languages (e.g. Extremeño, Aragonese), or just SEO lists consisting of e.g. song names not in Asturian. The proportions of unnatural texts vary a lot from language to language. Annotators report the following major types of unnaturalness: lists of services and goods, commercial ads with varying degrees of grammaticality, traces of Wikipedia markup, documents consisting mostly of menus, and boilerplate missed by boilerplate removal.

[33] Since the sample is stratified by group and the crawls from these groups give about 52% of all texts in our dataset, one should carefully interpret these statistics in the context of the full dataset.

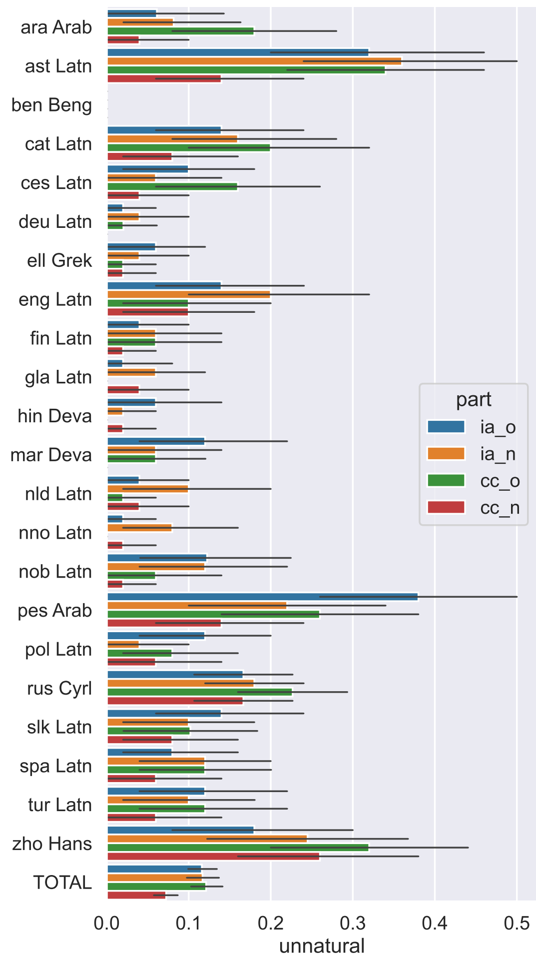

Figure 9: Proportions of unnatural texts among the cleaned texts extracted from four selected groups of crawls, according to manual inspection of a sample. Error bars correspond to the 95% confidence intervals.

Figure 9 shows proportions of unnatural texts for each language and group of crawls. Looking at individual languages, for most of them the group of new CC crawls give a much lower proportion of unnatural texts than other groups. However, since only 50 documents were labelled from each group and language, the confidence intervals are large and statistically significant conclusions cannot be made for each individual language. However, when annotations for all languages are combined (denoted as TOTAL on the figure) it becomes clear that for a random language (among those annotated) a random document has about 2x lower probability to be unnatural if it comes from the group of newer CC crawls compared to older CC crawls or any of two groups of IA crawls. For the proportions of pornographic content and documents misclassified by LID we did not observe any consistent differences for different groups of crawls.

I Model training and evaluation

I.1 Corpora comparison: English

Pretraining We fully replicated the original FineWeb training and evaluation setup by Penedo et al. (2024a), with the same architecture and pretraining settings (1.71B parameters, Llama architecture with a sequence length of 2048 tokens, GPT 2 tokenizer, and a global batch size of ~2 million tokens). We train 4 models that are differentiated only by training data, and evaluate their performance at different stages of model training. Each model is trained on 100 billion tokens, randomly sampled from the following datasets:

- English HPLT v2 data, cleaned
- English HPLT v2 data, deduplicated
- English HPLT v1.2 (de Gibert et al., 2024)
- FineWeb dataset (Penedo et al., 2024a)

We use NVIDIA's `Megatron-LM` (`https://github.com/NVIDIA/Megatron-LM`) training framework instead of HuggingFace's `nanotron` (`https://github.com/huggingface/nanotron`) framework used by Penedo et al. (2024a). Each model is trained on the LUMI supercomputer with 16 nodes, each with 4 AMD MI250x GPUs with dual-GCD (graphics compute die) design, amounting to 8 logical devices. In total, we used 128 devices and a single 64-core CPU for approximately 84 hours, totalling 11 008 GPU hours per model.

Evaluation Evaluation is performed using HuggingFace's `LightEval` tool (Fourrier et al., 2023) on the tasks listed below. Results per task are presented in Figure 10.

- **HellaSwag**: a dataset to evaluate commonsense reasoning. Its questions are designed to be trivial for humans but challenging for LLMs (Zellers et al., 2019).
- **PIQA**: a dataset focusing on reasoning with multiple-choice questions about physical interactions, evaluating the LLM's understanding of how different objects are used in various situations (Bisk et al., 2020).
- **OpenBookQA**: a dataset consisting of multiple-choice questions which require understanding concepts and their relations, benchmarking the complex reasoning and inference performance of the LLM (Mihaylov et al., 2018).
- **ARC Easy and ARC Challenge**: both parts of the AI2 Reasoning Challenge dataset, containing easier and more complex questions to test the LLM's reasoning skills (Clark et al., 2018).

I.2 Corpora comparison: Norwegian

Pretraining We mirrored the pretraining setup used for the English ablation studies in Appendix I.1, except for two details: 1) we trained a new tokenizer specifically for Norwegian, using a single tokenizer for all experiments trained on equal number of samples from all ablated corpora using the `tokenizers` library; 2) we pretrained the models for 30B tokens (roughly corresponding to 1 epoch on most of the ablated corpora) instead of 100B, mirroring the multilingual experiments for FineTasks (Kydlíček et al., 2024).

We compared five different filtered corpora that support Norwegian. Most of these discriminate between two written variants of Norwegian – Bokmål and Nynorsk – in those cases, we simply concatenate these subcorpora. The ablated corpora are:

- Norwegian HPLT v2 data, cleaned;
- Norwegian CulturaX (Nguyen et al., 2024a);
- Norwegian HPLT v1.2 (de Gibert et al., 2024);
- Norwegian FineWeb-2 (Penedo et al., 2024b);
- Norwegian mC4 (Xue et al., 2021).

The pretraining code is built on the Megatron-DeepSpeed framework (Smith et al., 2022). All models were trained on the LUMI supercomputer using 32 compute nodes, each with 4 AMD MI250x GPUs. The full pretraining run of each model took approximately 15 hours (wall-clock time), or 1 920 GPU-hours ($15 \times 32 \times 4$ hours), respectively.

Evaluation Evaluation is performed using `NorEval` (Mikhailov et al., 2025a), an open-source benchmark for Norwegian built upon LM Evaluation Harness (Gao et al., 2024). We consider the following ten multiple-choice QA, generative QA, sentence completion, and sentence pair ranking tasks that target different aspects of the model understanding and generation abilities in Norwegian Bokmål and Nynorsk:

- **Commonsense reasoning:** performing logical and commonsense reasoning (NorCommonsenseQA, Mikhailov et al., 2025b).
- **Norwegian-specific & world knowledge:** answering questions about facts and Norwegian culture (NorOpenBookQA and NRK-Quiz-QA, Mikhailov et al., 2025b).
- **Norwegian language knowledge:** understanding Norwegian punctuation rules (NCB,

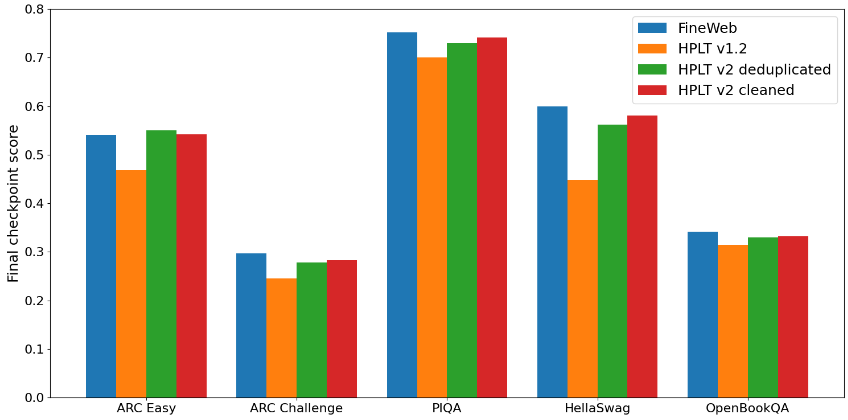

Figure 10: Final checkpoint scores of the English models trained on the datasets shown, grouped based on the benchmarks conducted. The models perform quite similarly with the exception of the model trained on the HPLT v1.2 dataset, the scores of which are noticeably lower.

Mikhailov et al., 2025a)[34] and idioms (NorIdiom, Mikhailov et al., 2025a).[35]

- **Machine reading comprehension:** understanding a given text and extracting an answer from it (NorQuAD, Ivanova et al., 2023).

We aim to find tasks that provide a reliable signal during pretraining. We evaluate the models in a zero-shot regime at regular checkpoint intervals (approx. 1B tokens) on all tasks. Next, we discard tasks that provide a low signal based on two criteria (Penedo et al., 2024b):

- **Monotonicity:** the Spearman rank correlation between the number of steps and the target performance score is at least 0.5 across all model checkpoints.
- **Non-random performance:** the difference between the random baseline (zero for generative tasks, one divided by the number of answer choices for multiple-choice tasks, and a coin flip probability for sentence pair ranking tasks) and the maximum score across all models is positive and satisfactory.

The filtering results in four datasets: NCB (accuracy), NRK-Quiz-QA Bokmål (accuracy), NorCommonsenseQA Bokmål (accuracy), and NorQuAD (F1-score). We aggregate the performance across the datasets using the average normalized score (Myrzakhan et al., 2024). We report the performance results for our 150 checkpoints in Figure 5 (see §6.2) and final checkpoint performance in Figure 11.

[34]https://huggingface.co/datasets/hcfa/ncb

[35]https://huggingface.co/datasets/Sprakbanken/Norwegian_idioms

J LTG-BERTs training and evaluation details

Following the HPLT v1.2,[36] we use UD treebanks of version 2.13[37] for most languages, except for Albanian and Georgian. These languages were not used in the HPLT v1.2 report due to missing training and development splits in UD 2.13. However, UD 2.15 does contain the required splits, and we use them. We do not evaluate NER on Maltese, since its WikiAnn training split contains only 100 samples. Table 15 shows detailed MLM evaluation results by language and task.

LTG-BERT architecture (Samuel et al., 2023) is a version of the original masked BERT model (Devlin et al., 2019). The differences include removing the next sentence prediction objective, swapping subword masking to span masking, and other minor architectural improvements. LTG-BERT was shown to perform well for small-sized training datasets (Samuel, 2023), which fits our evaluation setup. The models were trained with the same hyperparameters as in the aforementioned HPLT report.

We trained separate models for Bosnian and Croatian, in addition to the joint Bosnian-Croatian model. Since the UD does not provide Bosnian treebanks, we evaluated all three models on the Croatian datasets. We did not include Serbian, because it uses the Cyrillic writing system in HPLT v2, while UD features Serbian data only in Latin. Exploring whether mixing the scripts still improves the

[36]https://hplt-project.org/HPLT_D4_1___First_language_models_trained.pdf

[37]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5287

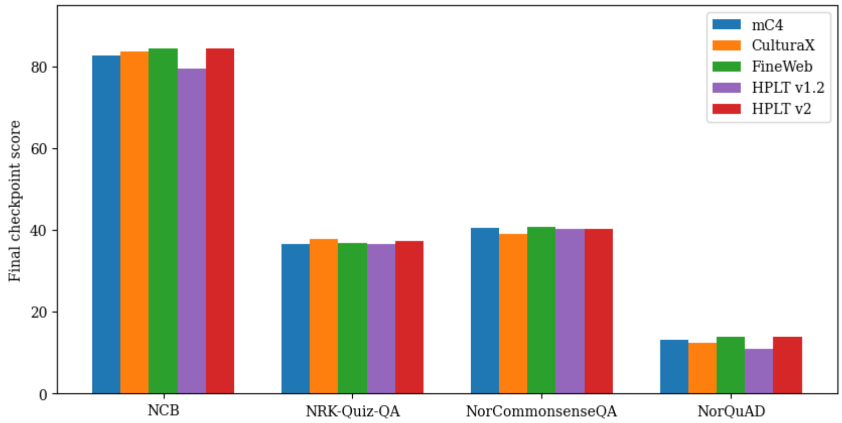

Figure 11: Final checkpoint scores of the Norwegian models trained on the datasets shown, grouped based on the evaluation datasets in `NorEval`. The models perform quite similarly with the exception of the model trained on the HPLT v1.2 dataset, the NorQuAD and NCB scores of which are generally lower.

results is left for future work. It is difficult to give any clear recommendations on which of the three models to use for practical tasks, since all of them yield satisfactory evaluation results (ranking varies from task to task).

LTG-BERT models were trained for 31 250 steps on 16 compute nodes with 4 physical AMD INSTINCT MI250x GPUs each for approximately 9.8 hours. Sharding, training a tokenizer and tokenizing for larger languages required up to 3.5, 0.5 and 1 hours correspondingly on 7 AMD EPYC 7763 CPUs (these numbers are estimated from the processing of English, the largest data subset in HPLT v2. The processing time of different languages may vary, for instance, languages without whitespace separation between words require an additional pretokenizing step). UD fine-tuning and NER fine-tuning required 1.1 hours and 8 minutes correspondingly on 1 GPU (estimated for English).

K Full Results for Translation Models Built on Parallel Data

We compare models trained on HPLT v2, Tatoeba (Tiedemann, 2012, 2020), and the combination of the two datasets. The language selection is the intersection of the languages covered by both datasets. We evaluate the models on the FLORES-200 evaluation benchmark (NLLB Team et al., 2024) using SacreBLEU implementation of BLEU[38] and chrF++[39] metrics (Post, 2018) and COMET-22-DA (Rei et al., 2022).

Tables 16 and 17 present the full results of the MT models for translation into English and from English respectively. For reference, we also include the performance of models trained on the HPLT v1.2 dataset, which shares the same underlying extraction pipeline. Note that we did not perform any language-specific hyper-parameter tuning which possibly led to low scores for a few model instances.

[38] `nrefs:1|case:mixed|eff:no|smooth:exp|version:2.5.1`, and where applicable, `tok:ja-mecab`, `tok:ko-mecab`, or `tok:13a`

[39] `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`

Language	POS tags				Lemmas				Dependency parsing				NER			
	mBERT	XLM-R	HPLT v1.2	HPLT v2	mBERT	XLM-R	HPLT v1.2	HPLT v2	mBERT	XLM-R	HPLT v1.2	HPLT v2	mBERT	XLM-R	HPLT v1.2	HPLT v2
als_Latn	59.1	61.6	64.0	**64.5**	**78.2**	75.0	76.3	77.2	**33.1**	29.3	25.3	24.7	92.3	92.9	92.4	**93.9**
bel_Cyrl	94.1	94.6	95.5	**95.7**	93.2	93.8	93.8	**97.1**	88.1	89.9	91.1	**91.7**	91.7	90.3	90.1	**92.8**
bos_Latn	95.5	96.2	96.4	**96.6**	97.2	**97.4**	97.2	97.1	90.2	91.3	91.3	**91.7**	91.5	91.6	89.3	**92.8**
hrv_Latn	95.5	96.2	96.4	**96.8**	97.2	**97.4**	97.2	97.1	90.2	91.3	91.3	**91.6**	91.5	91.6	89.3	**92.5**
bul_Cyrl	97.0	97.5	97.8	**97.9**	97.5	**97.7**	97.3	97.3	92.7	94.4	94.0	**94.5**	92.2	92.2	91.5	**93.0**
cat_Latn	97.1	97.2	97.4	**97.5**	99.4	99.4	99.4	**97.5**	93.6	94.1	94.4	**99.4**	92.1	91.0	90.1	**94.5**
ces_Latn	97.8	98.0	98.3	**98.4**	99.3	99.3	99.4	99.4	93.5	94.2	94.4	**94.6**	91.2	91.2	89.0	**91.8**
cym_Latn	87.2	88.3	**89.2**	89.0	**94.6**	94.4	93.7	92.3	80.8	**82.8**	82.3	**82.8**	92.5	90.0	89.4	**93.4**
dan_Latn	96.7	97.8	97.8	**97.9**	97.2	**97.6**	97.1	97.1	86.7	89.1	88.8	**89.5**	91.2	91.6	90.3	92.0
deu_Latn	88.8	89.4	80.7	**89.9**	97.6	**97.7**	95.5	97.5	84.6	87.1	76.4	**87.6**	**89.4**	87.7	64.1	89.2
ell_Grek	94.6	95.7	96.1	**96.2**	94.6	**94.7**	94.1	94.1	91.7	**93.5**	92.2	93.2	90.2	90.7	90.2	**92.6**
eng_Latn	96.1	96.8	96.7	**97.0**	97.8	98.0	97.9	**98.1**	91.3	92.6	92.2	**93.0**	2.2	81.1	81.0	**82.7**
spa_Latn	95.7	95.9	96.0	**96.2**	**99.4**	**99.4**	**99.4**	99.4	92.3	93.0	93.1	**93.4**	**90.9**	89.9	89.6	90.8
est_Latn	96.0	96.6	**97.1**	**97.1**	94.8	95.0	**95.2**	**95.2**	88.1	89.7	90.8	**91.0**	91.8	90.4	89.6	**93.0**
eus_Latn	91.0	91.4	**92.3**	**92.3**	95.7	95.9	**96.0**	95.9	85.3	87.3	88.1	**88.2**	91.3	90.7	89.8	**92.9**
pes_Arab	95.9	96.3	**96.4**	96.3	99.1	99.4	99.4	**99.5**	92.7	93.8	93.9	**94.1**	92.0	92.9	91.8	**93.9**
fin_Latn	95.1	96.4	96.8	**97.0**	90.6	91.5	**91.6**	91.4	90.2	93.0	93.3	**94.0**	90.2	90.0	89.2	**91.6**
fra_Latn	97.8	**98.1**	**98.1**	98.0	98.6	**98.8**	93.8	98.6	93.8	94.4	94.5	**94.8**	**90.5**	88.7	87.2	90.0
gle_Latn	86.5	87.1	88.7	**89.3**	95.5	95.8	**96.1**	95.6	81.3	82.7	83.4	**84.3**	**80.8**	78.0	55.9	78.2
glg_Latn	96.9	**97.1**	**97.1**	97.0	**98.3**	**98.3**	98.2	98.0	82.3	**82.6**	82.3	82.2	92.5	93.3	91.1	**94.1**
heb_Hebr	95.6	96.1	96.5	**96.7**	97.0	**97.2**	97.1	**97.2**	89.8	91.6	91.0	**91.9**	2.6	84.2	88.4	**89.3**
hin_Deva	92.4	93.3	93.6	**93.7**	98.9	**99.0**	**99.0**	**99.0**	92.6	93.3	93.5	**93.6**	88.6	88.0	84.3	**89.5**
hrv_Latn	95.5	96.2	96.4	**96.7**	97.2	**97.4**	97.2	97.2	90.2	91.3	91.3	**91.8**	91.5	91.6	89.3	**92.0**
hun_Latn	93.0	**94.3**	93.0	94.1	93.0	**94.3**	93.0	92.3	84.3	**86.7**	82.4	86.1	92.2	91.9	92.8	**93.1**
hye_Armn	88.7	91.2	**92.7**	**92.7**	94.4	**94.9**	93.9	94.7	80.4	85.3	84.1	**86.8**	95.7	95.3	94.8	**95.9**
ind_Latn	89.5	**89.8**	89.6	89.1	98.2	**98.3**	98.0	97.5	82.4	**82.7**	81.7	81.8	91.3	91.6	89.1	**92.0**
isl_Latn	87.7	88.1	88.6	**88.7**	96.2	96.4	**96.5**	96.4	85.2	86.6	86.9	**87.4**	**81.7**	63.9	55.9	78.3
ita_Latn	98.0	98.0	98.1	**98.3**	98.6	98.7	**98.8**	98.7	94.1	94.4	94.6	**95.1**	90.5	89.7	87.8	**91.2**
jpn_Jpan	97.5	97.7	**97.8**	**97.8**	98.3	98.3	98.3	**98.4**	94.1	94.6	94.6	**94.8**	66.5	65.9	**67.4**	67.2
kat_Geor	91.3	**92.6**	92.4	92.4	**92.8**	93.7	92.5	92.5	79.5	80.9	80.8	**81.3**	87.2	4.7	89.6	**90.7**
kor_Hang	88.6	89.7	89.9	**90.1**	94.0	94.3	**94.4**	94.4	88.0	89.0	89.4	**89.7**	87.8	87.0	88.3	**89.3**
lvs_Latn	91.6	92.8	92.4	**93.6**	96.9	91.6	96.8	**97.7**	88.8	90.9	90.9	**92.1**	93.2	92.6	90.7	**93.9**
lit_Latn	87.7	91.9	92.0	**92.5**	90.2	**91.6**	91.5	91.2	79.3	85.7	84.9	**86.8**	89.1	89.3	87.0	**91.0**
ltz_Latn	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.2
mkd_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.6
mlt_Latn	94.7	94.5	97.0	**97.7**	**100.0**	**100.0**	**100.0**	**100.0**	78.2	78.5	83.2	**87.2**	-	-	-	-
nob_Latn	97.0	97.4	**97.6**	97.5	98.5	**98.8**	**98.8**	98.7	93.2	94.3	94.5	**94.7**	91.9	92.6	91.1	**93.2**
nld_Latn	96.2	96.9	97.1	**97.2**	94.1	**94.7**	94.4	94.1	91.6	92.9	93.8	**94.1**	**91.7**	90.4	88.6	91.0
nno_Latn	96.6	97.0	97.7	**97.8**	98.2	98.4	**98.5**	**98.5**	92.9	93.9	94.6	**95.0**	**95.8**	93.6	93.2	95.5
pol_Latn	95.6	95.5	96.9	**97.2**	97.8	**98.2**	**98.2**	**98.2**	93.7	95.2	95.3	**95.6**	12.9	88.8	**89.7**	89.6
por_Latn	93.6	94.0	**94.1**	**94.1**	98.1	**98.3**	**98.3**	98.2	83.4	84.5	84.9	**85.3**	91.2	90.3	88.0	**91.5**
ron_Latn	97.3	97.6	97.7	**97.9**	97.7	**97.9**	97.8	97.8	89.5	91.0	90.6	**91.6**	**94.5**	93.6	91.2	93.6
rus_Cyrl	93.8	94.4	94.5	**94.7**	98.3	98.5	**98.6**	**98.6**	92.6	93.4	93.6	**93.8**	88.0	86.9	85.6	**89.0**
slk_Latn	89.1	97.6	**98.1**	91.9	95.7	**96.1**	95.6	95.5	92.9	94.4	93.8	**95.0**	93.2	92.9	91.2	**93.3**
slv_Latn	96.7	97.6	98.1	**98.2**	98.5	**98.7**	98.6	**98.7**	93.4	94.7	94.8	**95.3**	93.4	93.1	93.6	**94.2**
srp_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	91.6	92.4	-	**93.4**
swe_Latn	96.5	**97.4**	**97.4**	97.3	97.3	**97.6**	97.1	97.0	89.4	**92.1**	90.8	91.7	94.3	**94.5**	93.5	94.4
tat_Cyrl	-	-	-	-	-	-	-	-	-	-	-	-	**89.7**	80.6	82.9	84.0
tur_Latn	90.4	91.0	**91.5**	91.4	91.1	91.3	**91.9**	91.4	70.9	73.0	73.6	**74.6**	92.2	92.0	90.8	**92.5**
ukr_Cyrl	93.1	94.7	72.9	**95.3**	87.0	**97.2**	87.0	97.0	89.4	91.8	61.3	**92.1**	92.0	91.7	77.5	**92.8**
vie_Latn	89.8	**92.1**	91.8	**92.1**	**99.9**	**99.9**	**99.9**	**99.9**	66.5	**70.3**	68.0	**70.3**	**91.9**	90.6	89.2	90.3
zho_Hans	96.2	**96.3**	96.0	96.0	**99.9**	**99.9**	**99.9**	**99.9**	86.1	**86.9**	84.6	85.6	0.1	**76.5**	75.5	74.5

Table 15: Results of monolingual masked language models trained on the HPLT v2 datasets compared to the baselines on part-of-speech (POS) tagging, lemmatization, dependency parsing and named entity recognition. For POS tagging, we evaluate the AllTags performance, which is the exact match accuracy of the UPOS, XPOS and UFeats UDtags. For dependency parsing, we report LAS, and for lemmatization accuracy.

	HPLT v1.2			HPLT v2			Tatoeba (OPUS)			HPLT v2+OPUS		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
en-af				39.2	64.5	0.8398	38.3	63.6	0.8397	38.8	63.8	0.8409
en-ar				26.6	55.0	0.8442						
en-az				12.2	41.0	0.8128	11.3	38.7	0.8074	11.5	38.7	0.8011
en-be				11.6	39.0	0.7756	11.2	37.4	0.7767	11.8	38.2	0.7800
en-bg				38.0	62.4	0.8680	0.9	14.5	0.6774	30.0	51.9	0.8122
en-bn				16.0	45.2	0.8109	16.6	45.9	0.8282	16.8	46.1	0.8275
en-bs	4.7	26.0	0.4314	26.8	53.9	0.8672						
en-ca	38.4	61.7	0.8461	37.8	61.0	0.8334	39.8	62.2	0.8440	39.5	62.1	0.8425
en-cy				50.4	69.9	0.8611	47.7	67.6	0.8536	48.4	67.8	0.8506
en-eo				27.2	54.7	0.8264	25.5	54.1	0.8500	25.9	54.4	0.8523
en-et	23.7	53.4	0.8664	24.5	53.8	0.8684	24.5	53.3	0.8600	24.4	53.3	0.8578
en-eu	12.1	43.4	0.7674	16.5	49.5	0.8215	14.9	47.2	0.8098	14.8	47.1	0.8122
en-fa				21.5	47.5	0.7947	23.4	50.0	0.8336	23.6	50.0	0.8349
en-fi				21.3	51.1	0.8709	21.3	51.2	0.8725	22.1	51.6	0.8752
en-ga	27.3	52.6	0.7561	29.0	53.9	0.7543	30.2	53.9	0.7715	30.8	54.6	0.7717
en-gl	27.9	54.0	0.8033	30.0	55.7	0.8179	31.4	56.1	0.8302	31.4	56.1	0.8264
en-gu				19.3	46.5	0.8066	22.5	49.9	0.8518	22.6	49.9	0.8479
en-he				28.1	54.0	0.8320	29.7	55.9	0.8532	29.6	55.4	0.8503
en-hi	32.8	55.5	0.7621	32.0	54.6	0.7612	33.1	55.5	0.7728	32.5	54.9	0.7658
en-hr				27.5	53.7	0.8504						
en-is	20.6	45.1	0.7651	22.2	47.1	0.7766	22.8	47.1	0.7800	23.1	47.5	0.7859
en-ja				27.0	26.6	0.8244	29.9	30.2	0.8640	29.6	29.9	0.8633
en-kk				21.0	51.4	0.8651	16.5	45.1	0.8315	16.9	45.3	0.8347
en-kn				13.8	43.5	0.7746	19.5	50.8	0.8348	19.2	51.1	0.8369
en-ko				25.0	31.2	0.8268	26.6	32.2	0.8424	26.5	32.0	0.8402
en-lt				22.9	50.9	0.8214	25.0	52.7	0.8486	24.4	52.5	0.8417
en-lv				26.8	53.1	0.8214	23.9	50.0	0.7898	24.3	50.6	0.7891
en-mk				32.2	58.6	0.8478	33.9	60.1	0.8718	34.1	60.1	0.8660
en-ml				0.6	20.2	0.5753	14.4	47.9	0.8438	14.6	48.0	0.8427
en-mr				11.0	37.9	0.6086	13.9	42.3	0.6808	14.2	42.4	0.6792
en-ms				38.3	63.9	0.8580	24.4	52.6	0.8534	25.1	53.2	0.8540
en-mt				34.7	63.2	0.7072	35.8	64.1	0.7102	36.0	64.4	0.7111
en-nb				33.0	58.2	0.8661						
en-ne				0.9	20.9	0.4880	12.9	42.8	0.7404	12.8	43.1	0.7386
en-nn				22.8	47.1	0.7766						
en-si				1.2	18.6	0.6289	13.1	41.0	0.8542	13.2	41.2	0.8548
en-sk				29.3	54.0	0.8279	29.3	53.9	0.8353	29.9	54.5	0.8423
en-sl				26.8	52.2	0.8295	26.5	52.0	0.8339	27.5	52.6	0.8414
en-sq	27.8	54.6	0.8509	27.7	54.2	0.8398	29.9	55.8	0.8659	29.3	55.3	0.8600
en-sr				32.2	57.5	0.8512						
en-sw	28.4	54.6	0.7743	32.5	58.2	0.7964	31.2	57.0	0.8058	31.3	57.0	0.8031
en-ta				14.1	46.7	0.8139	16.7	50.3	0.8565	16.4	50.1	0.8553
en-te				20.2	51.3	0.8104	22.1	53.7	0.8378	22.7	53.9	0.8383
en-th				9.9	40.9	0.7977	8.1	40.6	0.8053	8.7	40.9	0.8053
en-tr				25.3	53.7	0.8368	27.8	56.4	0.8685	27.5	55.8	0.8638
en-uk				26.7	52.6	0.8457	27.2	53.4	0.8532	26.8	52.8	0.8471
en-ur				18.9	43.2	0.7548	19.3	44.0	0.7537	19.5	44.6	0.7584
en-uz				16.3	49.1	0.8397	15.9	47.3	0.8497	17.1	48.8	0.8532
en-vi				37.8	55.8	0.8358	39.3	57.1	0.8489	38.8	56.6	0.8451
en-xh				12.0	44.2	0.7323	0.0	3.5	0.2317	0.0	3.5	0.1647

Table 16: MT results (BLEU, chrf++, COMET22-DA) for models translating from English, trained on our HPLT v2, Tatoeba (OPUS), a combination of both, and the existing HPLT v1.2 (numbers reported where available).

	HPLT v1.2			HPLT v2			Tatoeba (OPUS)			HPLT v2+OPUS		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
az-en				18.5	47.1	0.8290	17.4	44.7	0.8039	18.6	46.2	0.8175
be-en				16.1	46.7	0.7886	14.8	44.7	0.7621	15.5	45.8	0.7743
bg-en				35.5	61.1	0.8556	7.4	32.5	0.5104	34.8	60.6	0.8524
bn-en				27.9	53.7	0.8468	28.4	53.7	0.8498	29.0	54.2	0.8523
bs-en	12.8	38.0	0.5882	35.2	59.9	0.8508						
ca-en	41.0	64.4	0.8676	39.2	63.1	0.8478	41.1	64.4	0.8580	40.3	63.9	0.8541
cy-en				51.5	70.7	0.8615	50.0	68.8	0.8456	50.6	69.1	0.8455
eo-en				35.5	59.2	0.8362	35.6	59.2	0.8437	35.6	59.3	0.8429
et-en	30.6	56.6	0.8611	30.3	55.8	0.8517	30.7	56.1	0.8510	30.7	55.6	0.8510
eu-en	19.4	45.7	0.7810	23.3	49.2	0.8121	22.2	47.5	0.8064	22.0	47.4	0.8042
fa-en				31.1	56.4	0.8447	33.7	58.2	0.8585	32.7	57.6	0.8546
fi-en				26.6	52.4	0.8442	26.6	51.9	0.8451	26.2	51.6	0.8421
ga-en	29.9	54.9	0.7653	34.1	58.8	0.8006	32.3	56.3	0.7754	33.1	57.7	0.7918
gl-en	31.4	57.2	0.8236	33.7	59.2	0.8374	34.5	59.1	0.8395	35.0	59.9	0.8441
gu-en				28.5	54.6	0.8475	32.0	57.0	0.8646	33.0	57.6	0.8667
he-en				38.2	62.2	0.8534	39.7	62.9	0.8622	40.4	63.6	0.8665
hi-en	35.2	59.9	0.8741	34.7	59.5	0.8701	35.8	60.1	0.8738	36.9	61.0	0.8773
hr-en				31.7	56.4	0.8389						
is-en	25.3	50.0	0.7815	29.0	53.4	0.8189	29.0	52.8	0.8163	28.7	52.8	0.8136
ja-en				19.9	46.8	0.8255	24.6	52.5	0.8628	23.6	50.6	0.8533
kk-en				27.0	53.4	0.8403	22.6	47.8	0.8003	22.6	47.7	0.7998
kn-en				3.8	24.5	0.6246	27.9	53.4	0.8391	27.4	53.2	0.8396
ko-en				24.1	51.3	0.8458	25.7	52.7	0.8586	25.8	52.7	0.8578
lt-en				26.5	51.9	0.8138	27.2	52.6	0.8224	27.0	52.2	0.8201
lv-en				29.3	56.0	0.8368	25.0	50.9	0.7862	26.6	53.3	0.8113
mk-en				37.0	61.4	0.8522	38.2	62.0	0.8558	38.8	62.6	0.8596
ml-en				2.9	23.5	0.5978	26.4	51.7	0.8342	26.4	51.9	0.8363
mr-en				23.8	49.8	0.8063	26.1	51.9	0.8299	26.9	52.2	0.8320
ms-en				37.2	61.3	0.8561	38.5	61.8	0.8579	38.0	61.7	0.8583
mt-en				45.0	68.4	0.7777	47.1	68.5	0.7892	47.5	68.9	0.7895
nb-en				37.3	61.0	0.8540						
ne-en				22.1	47.9	0.7929	26.6	52.0	0.8408	27.7	52.8	0.8436
nn-en				32.9	55.8	0.7945						
si-en				3.0	24.2	0.5979	26.0	51.2	0.8381	26.9	51.9	0.8418
sk-en				31.6	58.1	0.8456	32.7	58.6	0.8486	33.2	59.0	0.8535
sl-en				29.2	55.0	0.8371	28.7	54.4	0.8345	29.7	55.6	0.8402
sq-en	31.7	58.3	0.8468	32.1	58.6	0.8453	33.7	58.8	0.8448	34.8	59.8	0.8534
sr-en				37.4	62.7	0.8544						
sw-en	27.2	51.0	0.7542	35.3	57.8	0.8086	34.3	56.3	0.7979	33.5	55.6	0.7932
ta-en				24.0	49.6	0.8068	23.4	49.0	0.8139	24.3	49.5	0.8154
te-en				30.2	55.3	0.8328	31.5	55.9	0.8438	31.9	56.4	0.8446
th-en				24.9	52.3	0.8452	22.9	51.0	0.8381	23.7	51.7	0.8410
tr-en				29.5	54.9	0.8392	32.2	57.3	0.8622	32.7	57.4	0.8602
uk-en				33.1	58.7	0.8444	33.4	59.2	0.8470	33.9	59.6	0.8478
ur-en				26.3	52.1	0.8138	26.2	50.9	0.8097	27.4	52.0	0.8144
uz-en				24.8	51.5	0.8110	23.6	48.6	0.7990	24.8	50.0	0.8064
vi-en				32.0	56.4	0.8514	33.5	57.9	0.8602	32.9	57.2	0.8543
xh-en				22.1	45.5	0.6703	0.1	11.8	0.2871	0.2	13.1	0.2628

Table 17: MT results (BLEU, chrf++, COMET22-DA) for models translating into English, trained on our HPLT v2, Tatoeba (OPUS), a combination of both, and the existing HPLT v1.2 (numbers reported where available).