HPLT's Second Data Release

Nikolay Arefyev[2], Mikko Aulamo[3], Marta Bañón[4], Laurie Burchell[1], Pinzhen Chen[1], Mariia Fedorova[2], Ona de Gibert[3], Liane Guillou[1], Barry Haddow[1], Jan Hajič[5], Jindřich Helcl[5], Erik Henriksson[6], Andrey Kutuzov[2], Veronika Laippala[6], Bhavitvya Malik[1], Farrokh Mehryary[6], Vladislav Mikhailov[2], Amanda Myntti[6], Dayyán O'Brien[1], Stephan Oepen[2], Sampo Pyysalo[6], Gema Ramírez-Sánchez[4], David Samuel[2], Pavel Stepachev[1], Jörg Tiedemann[3], Dušan Variš[5], Jaume Zaragoza-Bernabeu[4]

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

Contact: https://hplt-project.org

Abstract

We describe the progress of the High Performance Language Technologies (HPLT) project, a 3-year EU-funded project that started in September 2022 with two main objectives: derive monotexts and bitexts for multiple languages from web crawls at massive scale and use them to build efficient machine translation models and language models. We focus on the up-to-date results on the release of free text *datasets* derived from web crawls, one of the central objectives of the project. The second release used a revised processing pipeline, and an enlarged set of input crawls. From 4.5 petabytes of web crawls we extracted 7.6T tokens of monolingual text in 193 languages, plus 380 million parallel sentences in 51 language pairs. We also release MultiHPLT, a cross-combination of the parallel data, which produces 1,275 pairs, and the containing documents for all parallel sentences in order to enable research in document-level MT. We report changes in the pipeline, analysis and evaluation results for the second parallel data release based on machine translation systems. All datasets are released under the CC0 licence.

1 Introduction

The HPLT project runs from 2022 to 2025, and focuses on the processing petabytes of natural language data and large-scale model training. The consortium is made of eight partners: Charles University in Prague (coordinator), University of Edinburgh, University of Helsinki, University of Oslo, University of Turku, Prompsit Language Engineering, and CESNET and Sigma2 HPC centres.

Following the previous release at the end of 2023 (de Gibert et al., 2024), the project has recently completed the release of a new massive multilingual dataset for both monolingual and parallel data along with improved pipelines and tools extensively described in (Burchell et al., 2025).

2 Second Data Release

Datasets The second release includes data processed originally from 4.5 petabytes of the Internet Archive and CommonCrawl to create monolingual and parallel corpora. It is released under the permissive CC0 licence[1] through our project website[2], OPUS[3][4] and Hugging Face[5]. The updated pipelines and open-source tools to produce this release are on GitHub.[6] The monolingual data extends to 193 languages and contains roughly 7.6 trillion space-separated tokens after deduplication and filtering. The parallel data includes 51 language pairs, with roughly 6.7 billion tokens computed on the English side and 380 million sentence pairs. The bonus multi-parallel dataset, pivoted through English, contains 1,275 language pairs.

Changes in the parallel data pipeline The second release of HPLT data introduces important changes in the pipeline. Parallel data is now derived from the clean and deduplicated documents from the monolingual release instead of the raw data. The text extraction pipeline uses Trafilatura (Barbaresi, 2021), which results in more efficient boilerplate removal. Language identification uses a refined version of OpenLID (Burchell et al., 2023) instead of CLD2. Deduplication and filtering of adult content and non-compliant `robots.txt` web documents happens before executing the parallel data processing. A multilingual Bicleaner AI[7] model replaces the pair-based ones used to annotate parallel sentences for translation likelihood.

[1]We do not own any of the text from which these text data have been extracted. We license the actual packaging of these text data under the CC0 licence ("no rights reserved").

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

[3]opus.nlpl.eu/HPLT.php

[4]https://opus.nlpl.eu/MultiHPLT/corpus/v2/MultiHPLT

[5]https://huggingface.co/datasets/HPLT/HPLT2.0_cleaned

[6]github.com/hplt-project

[7]https://tinyurl.com/3pxkcjf8

Proceedings of Machine Translation Summit XX Volume 2, pages 101–102, June 23-27, 2025

Parallel data stats and analysis The second release of the HPLT parallel data covering 51 language pairs contains 380,710,720 sentence pairs with 6,779,910,082 English words. Our selection of pairs avoided the top 20 highest resourced (according to OPUS) and focused on the next ranked languages, in order to maximise impact. The sizes of the different language pairs show significant variation, with a median of 3,927,371 sentence pairs. This range spans from 273,430 sentence pairs for English– Sinhala to 29,067,875 sentence pairs for English–Finnish, the largest parallel data set.

We get a 36% increase in the number of sentences compared to the first release for the 18 overlapping languages. During filtering, 40% of the parallel sentence pairs are eliminated and an additional 50% is removed due to deduplication. The final corpus shows a 70% decrease in sentence pairs relative to the raw data. This reduction is less significant than in the first release, which we assume is due to starting with cleaner monolingual text.

We inspect the data with the HPLT Analytics tool.[8] We find that small-sized datasets contain larger portions of Wikipedia and religious content while medium/large-sized ones contain high-portions of hotel booking and travel websites. Some popular domains include websites from gaming, software or e-commerce translated into a big number of languages, probably using MT. From the inspection of the most frequent n-grams, we find that they are very similar across all parallel datasets, especially among larger ones, frequently related to hotels and legal notices.

Extrinsic evaluation of the parallel data We train bidirectional MT models on the new released parallel data to extrinsically evaluate the performance of the released datasets. We compare models trained on only HPLT data for both releases and, additionally, models trained with HPLT data in combination with Tatoeba.[9]

We build and release MarianNMT compatible MT models for all bitexts in HPLT v2 using the same tooling as the one used in the previous release: OpusCleaner (data selection and cleaning), OpusTrainer (data scheduling and augmenting), and OpusPocus (training process management). These tools are fully described in the public deliverable of the HPLT project focused on pipelines and tools.[10]

Automatic metrics are computed on FLORES-200[11] for evaluation. Results computed on 10 out of the overlapping 18 language pairs between the first and the second release show gains in BLEU in favour of HPLT v2 MT models going into English with an average gain of 4.2 BLEU. From English, the average gain is 3.5 in BLEU, with 7 out of the 10 models being better with HPLT v2 data and the remaining 3 being on par between the first and second release. When combining HPLT v2 data and Tatoeba, MT models result in a 7% relative increase in BLEU for both translation directions.

Acknowledgment

This project has received funding from the European Union's Horizon Europe programme (GA No 101070350) and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (GA No 10052546). It has also been supported by the Czech MEYS project No. CZ.02.01.01/00/23_025/0008691 and Research Infrastructure project LM2023062.

References

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of ACL Demo*.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of ACL)*.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An expanded massive multilingual dataset for high-performance language technologies. *Preprint*, arXiv:2503.10267.

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 LREC-COLING*.

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

[9] https://github.com/Helsinki-NLP/Tatoeba-Challenge

[10] https://tinyurl.com/y6mc3sfk

[11] https://github.com/facebookresearch/flores