



High-Performance Language Technologies (HPLT): bulding LLMs and TMs in Europe

Jan Hajič

Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic

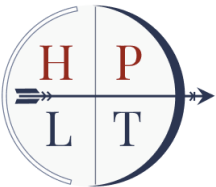




The HPLT project

- High-Performance Language Technology
- Horizon Europe DATA call, 2022-2025
- Goals
 - Collect large data from Internet Archive (San Francisco, CA, USA)
 - Approx. 12 PB
 - Extract text, clean, identify, deduplicate, pseudonymize, describe, ...
 - Both monolingual and bilingual (parallel) data (texts only)
 - Train language and translation models: 24 EU + min. 16 other
 - xBERTy, GPT-x, Transformer, future SoTA
 - make them openly available (OpusMT, Huggingface, possibly other repos)
 - Evaluate models – keep a dashboard
 - Demonstrate use of EU HPC Centres in a distributed manner
 - Huge compute demands: just for cleaning, 20 mil. CPU hours
 - Millions of GPU hours needed for LLMs building





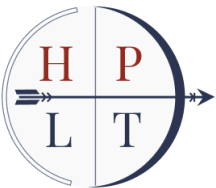
HPLT project partners

- Charles University
 - ÚFAL/LINDAT, Jan Hajič, Dušan Variš, Jindřich Helcl, Martin Popel, Pavel Straňák, Barbora Vidová Hladká)
 - coordinator
- University of Edinburgh (Scotland, UK, Barry Haddow / formerly Ken Heafield)
- University of Helsinki (Finland, Jorg Tiedemann, OpusMT)
- University of Turku (Finland, Sampo Pyysalo, Filip Ginter)
- University in Oslo (Norway, Stephan Oepen)
- Prompsit (Spain, Gema Ramirez)
- HPCs:
 - CESNET (Czechia, Luděk Matyska, David Antoš)
 - Sigma2 (Norway, Hans Eide)
 - Cooperation with LUMI, EuroHPC, Karolina (IT4Innovations), possibly others

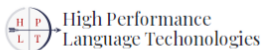




The HPLT project status: Data



- Almost 50% of the data has been developed from IA
 - Under review
 - Will be used for training
 - IA advanced
 - Will be used for training
- Data published
 - Plain text
 - 8.4 TB
 - Bilingual
 - <https://hpl.tugraz.at/>
 - Legal disclaimer
- Ongoing:



About Publications Deliverables Dashboards **Models** **Datasets**

HPLT Datasets v1.2

This version elaborates upon [version 1.0](#), which was made of mostly raw plain text data. The previous one, v1.1 has been deprecated.

What's new in v1.2

- We fixed a bug found in the de-duplication algorithm for monolingual data in version 1.1. Both monolingual and bilingual datasets are now correctly deduplicated. And we recovered back a lot of monolingual data!
- Further cleaning has been applied to monolingual datasets. Full documents have been filtered following 5 criteria: URL is in UT1 blacklist of adult sites, average words per segment is less than 5, it contains less than 200 characters or less than 5 segments (lines) and less than 20% of the segments in a document share the language identified at document level.
- The bilingual datasets have now been anonymized using a blend of regular expressions and NER on the English side as implemented in [BIROAMer](#).

For further information about how these datasets were produced or if you use them, please read and cite "A New Massive Multilingual Dataset for High-Performance Language Technologies".

Monolingual

Data release 1.2 (December 2023)

There are 75 languages in this release (22 TB of raw files, 11 TB of deduped files and 8.4 TB of clean files) provided as JSONL files compressed with zstd. For convenience, data is split into multiple shards, a few GB each. The number of shards per language depends on the size of the specific corpus.

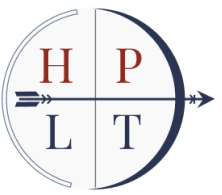
The format is JSONL, where each line is a valid JSON value and a full document with metadata. For example:

Bilingual

Data release 1.2 (December 2023)

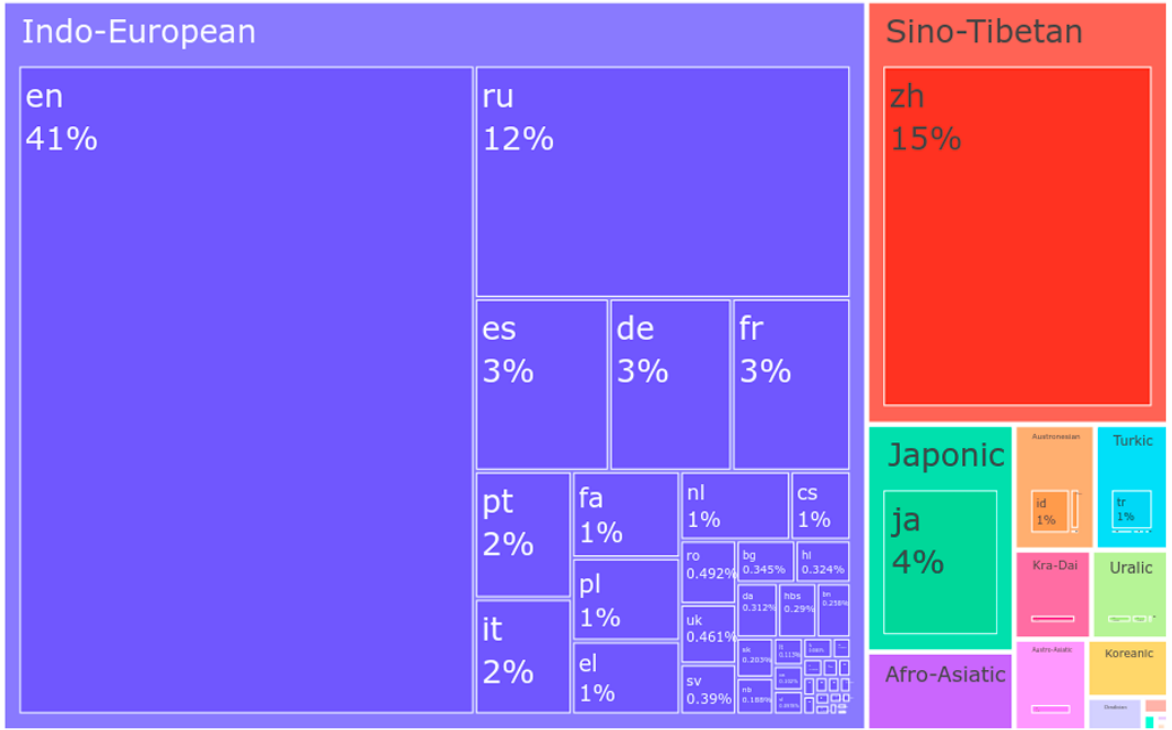
There are 18 language pairs in this release. The parallel corpus contains over 96 million clean and unique sentence pairs and covers over 1.4 billion English tokens. The corpora are provided in raw, TMX and TXT compressed formats. These corpora have been highly curated, de-duplicated and filtered using the full [Bitextor](#) pipeline. Besides this, an anonymized (ROAM) version of the TMX is also provided.





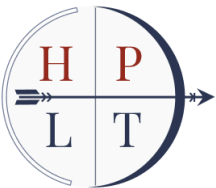
The HPLT project status: Data

- Almost 5 PB
 - Under spec
 - Will be collected
 - IA advantage
 - Whole
- Data published
 - Plain text, assigned,
 - 8.4 (54)
 - Bilingual
 - <https://hpl>
 - Legal constraints
- Ongoing: advertisement



Size distribution for the monolingual corpora, organized by language family and language.





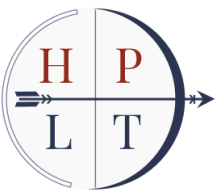
The HPLT project status: LLMs

- Encoder only (BERT)
 - 75 languages - 75 **BERT-xy** models, 123M parameters
 - Evaluates better than mBERT or XLM-R on same tasks
- Distribution
 - <https://hplt-project.org/models/llm>
 - <https://huggingface.co/HPLT>
 - intermediate checkpoints published as well
- Generative LLMs (Apache 2.0 lic.)
 - **FinGPT**: monolingual models for **Finnish** (completed), up to 13B
 - **NORA.LLM**: monolingual models for **Norwegian** (completed)
 - **Poro 34B**: model for **Finnish** and **English** (completed), incl. code
 - **Viking, Europa**: under construction – with Silo.AI, HF, others




The HPLT project status: TMs

- Translation models
 - Univ. of Helsinki in HPLT
 - Uses also previous resources collected in Opus MT projects
- 16 language pairs, 32 directions
 - 3 sources: OPUS only, HPLT only, HPLT+OPUS
- Distribution
 - HuggingFace, HPLT Github
 - Frameworks:
 - MarianNMT (adapted for AMD/LUMI) and transformers



The HPLT project status: TMs

- Trans
- Uni
- Use
- 16 lar
- 3 so
- Distric
- Hug
- Frar
- |



Find your corpora

Source language

Afar (source) ▾

Target language

Abkhazian (target) ▾

Q Search

An overview of the OPUS collection

Hug **1,210** CORPORA

Frar **45,945,946,108** TOTAL SENTENCE PAIRS

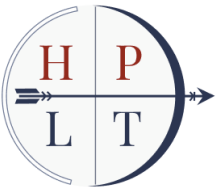
| **744** LANGUAGES AVAILABLE

THIS MAP DISPLAYS **10** CORPORA , WHICH MAKE UP A TOTAL **93.40%** OF THE ENTIRE **OPUS** COLLECTION

See next 10 →

Corpus	Sentences	% of OPUS
NLLB	13B	28.31
CCMatrix	11B	23.64
OpenSubtitles	8.5B	18.53
MultiCCAligned	2.2B	4.87840
ParaCrawl	1.5B	3.26229
DGT	1.1B	2.37845
XLEnt	883M	1.92148
MultiParaCrawl	789M	1.71653
LinguaTools-WikiTitles	487M	1.06082
CCAligned	439M	0.95442





HPLT data processing

- Monolingual / bilingual pipelines
 - **warc2text** tool (IA distributed in WARC packages)
 - **Monofixer** tool for fixing character encoding, removing HTML
 - **FastText** for language ID, plus **CLD2**
 - Fluency score by Knesser-Ney character language model
 - Packaging ("sharding") into equal-sized chunks (a few GBs)
- Bilingual
 - Follows after the monolingual pipeline runs
 - Separated into sentences, translated to English (MarianNMT / OpusMT data), determined similarity / match, cleaned, packaged
 - Using **Bitextor** pipeline (incl. **Bluealign**, **Bifixer**, **Bicleaner**)
- Run on Karolina, Sigma2 (CPU-based)





HPLT additional processing

- Going from v1.0 to v1.2
 - Deduplicated
 - Further cleaning
 - Filtered out documents based on the UT1 blacklist of adult sites
 - Filtered out segments with words per segment < 5, or 200 characters
 - Filtered out mixed language documents (20% of segments or less share the lang ID of document)
 - Parallel data (bitexts) anonymized using the **BiROAMer** tool
 - Mix of Named Entity Recognition and regular expressions

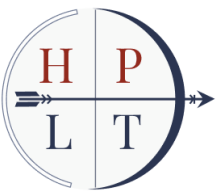




HPLT: data privacy issues (tech)

- Pseudonymization vs. anonymization
 - In text data, similar issue (vs. speech etc.)
 - Could still contain unanonymizable data in some text types
- Standard methodology
 - Named Entity Recognition
 - Removal (by regular expressions)
- Problems faced
 - Non-English languages
 - Scarcity or non-existence of NER training data, person names wrongly marked or mixed up
 - Circumventing: translate, solve NER, align and map back to original language
 - Identifying information outside of names
 - Place of birth, date of birth, IDs, ...
 - Some of it solved by regular expressions – problem of multilinguality, errors

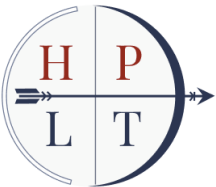




HPLT: data privacy issues (legal)

- Right to forget / removal
 - All data marked by a privacy clause that requires us to remove any offending data as reported by eligible users:
 - *Take down: We will comply to legitimate requests by removing the affected sources from the next release of the corpora.*
- Much bigger problem: copyright
 - As with all internet data today wrt Generative models
 - Texts least problematic
 - Specific evaluation of LLMs needed (all aspects)
 - Will be developed later in the project when LLMs are built





Thank you!

<https://hplt-project.org>

Twitter: [@hplt_eu](https://twitter.com/hplt_eu)

<https://ufal.mff.cuni.cz>

<https://lindat.cz>

<https://lindat.cz/services>

Twitter: [@LindatClariahCZ](https://twitter.com/LindatClariahCZ)

Twitter: [@ufal_cuni](https://twitter.com/ufal_cuni)

