EUROPEAN LANGUAGE EQUALITY

META-NET

META-FORUM 2023

HPLT: High Performance Language Technologies

Barry Haddow (University of Edinburgh)
bhaddow@ed.ac.uk

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

We propose a language data space that substantially lowers the three main barriers to training at scale: data gathering, compute, and reproducibility.

[HPLT Proposal]

1

Project Overview

What are we doing?

Ingest **Petabytes** of raw data to create:

→ Consistently formatted and curated **language data sets**
→ Efficient and high-quality **language** and **translation models**
→ Sustainable and reusable **workflows** using HPC

When?

- Project started in October 2022 (for 3 years)
- First data release due September 2023

HPLT in Numbers

7 petabytes of web data from the internet archive

5 petabytes of web data from commoncrawl

2.5 trillion words of monolingual text

600 unique corpora

~80 languages to cover

100s of efficient language and translation models

36 months to complete the project

8 consortium partners collaborating together

Monolingual text targets

Words	#Langs	Languages
> 1.5 trillion	1	en
> 100 billion	4	de, es, fr, ru
> 10 billion	21	ar, cs, da, el, fa, fi, he, hu, id, it, ja, ko, nb, nl, pl, pt, ro, sv, tr, uk, vi, zh
> 1 billion	19	az, be, bg, bn, ca, et, hi, hr, la, lt, lv, ms, sk, sl, sq, sr, ta, th, zh-Hant
> 100 million	31	af, cy, eo, eu, ga, gl, gu, hy, is, ka, kk, kn, ky, mk, ml, mn, mr, mt, my, ne, nn, pa, ps, si, so, sw, te, tl, tt, ur, uz

Data and Model Releases

September 23	March 24	September 25
Data	**Data & Models**	**Data & Models**

Corpora

- Monolingual corpora for 80+ languages; Bitexts for 150+ pairs

Language Models

- Decoder-only, Encoder-only, Enc-Dec – For all languages

Translation Models

- High-performance translation models for all language pairs

Tooling

Aim: Open-source tools for complete crawling and model-building pipeline. Dashboards for monitoring.

Bitextor

Automatically harvests parallel text from websites

OpusCleaner

Select, download and clean parallel and monolingual corpora.

OpusTrainer

Large-scale curriculum training for MT

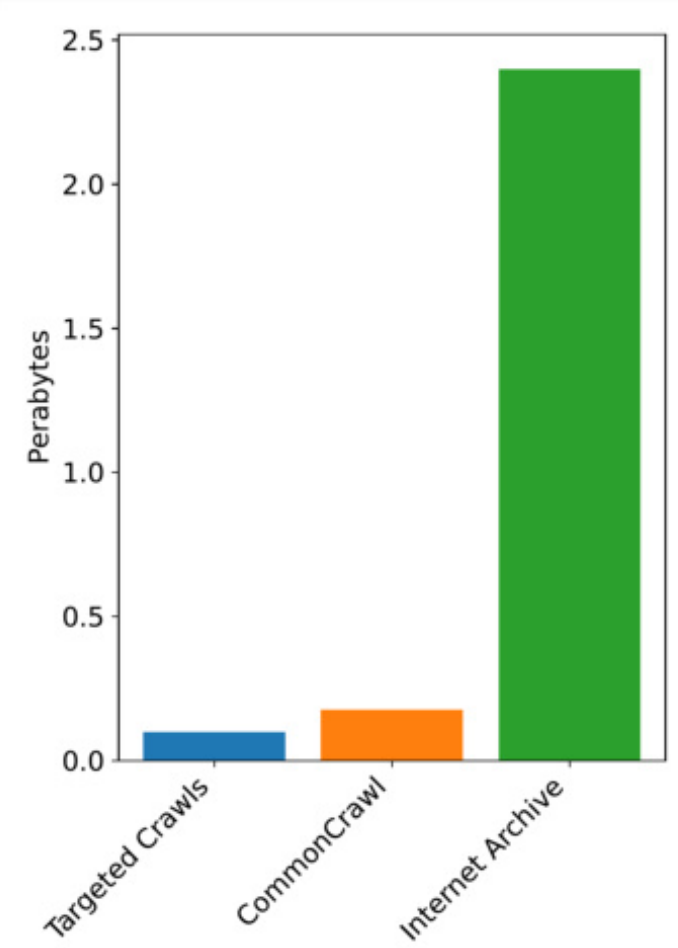
Megatron

LLM training – ported to AMD hardware

OpusAPI

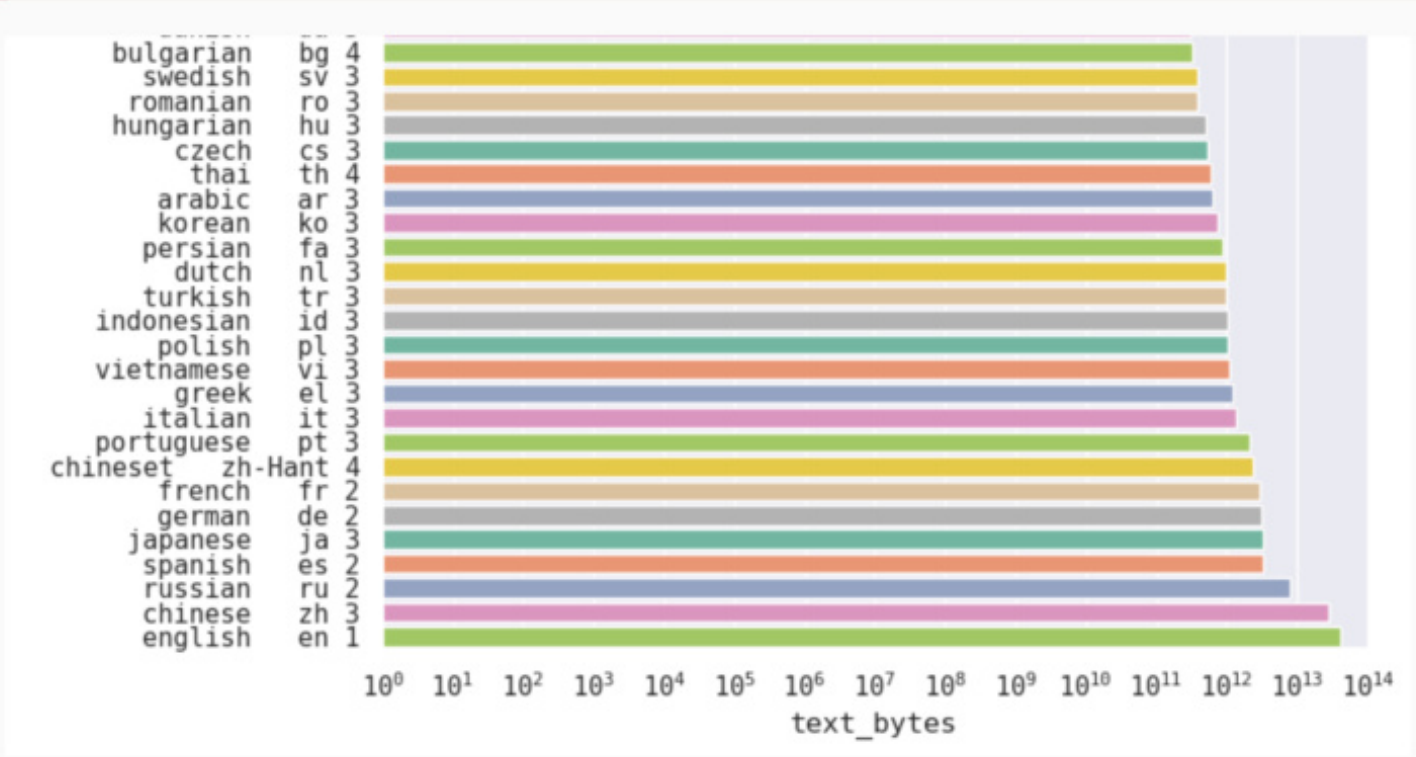
Access Opus corpora and models

Download Progress



- Total downloaded for 2023 release
- Compressed web archive (WARC)
- Targeted crawls from earlier projects
  - Mainly ParaCrawl
- CC segments from 2022
- IA from 5 “wide” segments

Preliminary Language Statistics



- Bytes of uncompressed text for top 25 languages
- Document-level langid with CLD2

Towards a first data release

Monolingual Text

- Aim for up to 95 languages
- Expect $\approx$ 90TB (compressed)
- Deduping will remove at least 1/3

Bitext

- Up to 46 langs (paired with English)
- Bitextor pipeline
- Fast marian models
- Running on LUMI

Contact

web: https://hplt-project.org twitter: @hplt_eu

Partners


CHARLES UNIVERSITY


UNIVERSITY OF OSLO


UNIVERSITY OF EDINBURGH


UNIVERSITY OF TURKU


UNIVERSITY OF HELSINKI


PROMPSIT


CESNET


SIGMA2

Acknowledgement

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant no 10052546].

ELE

EUROPEAN LANGUAGE EQUALITY

Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

Barry Haddow (University of Edinburgh)
bhaddow@ed.ac.uk

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu