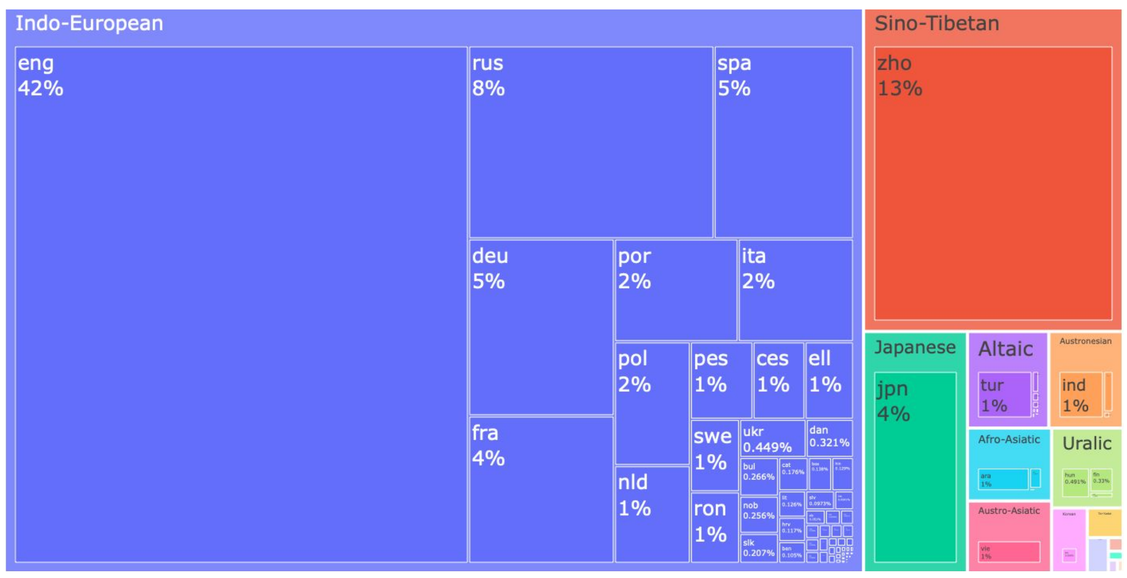
HPLT v2 is a massive open corpus of high-quality textual training data, covering 193 languages.

Background: Training state-of-the-art large language models requires **vast amounts of clean and diverse textual data.** However, there is a lack of the **open multilingual datasets** we need for research.

The HPLT v2 datasets

✨ **20TB** of clean data, permissive **CC0 license**

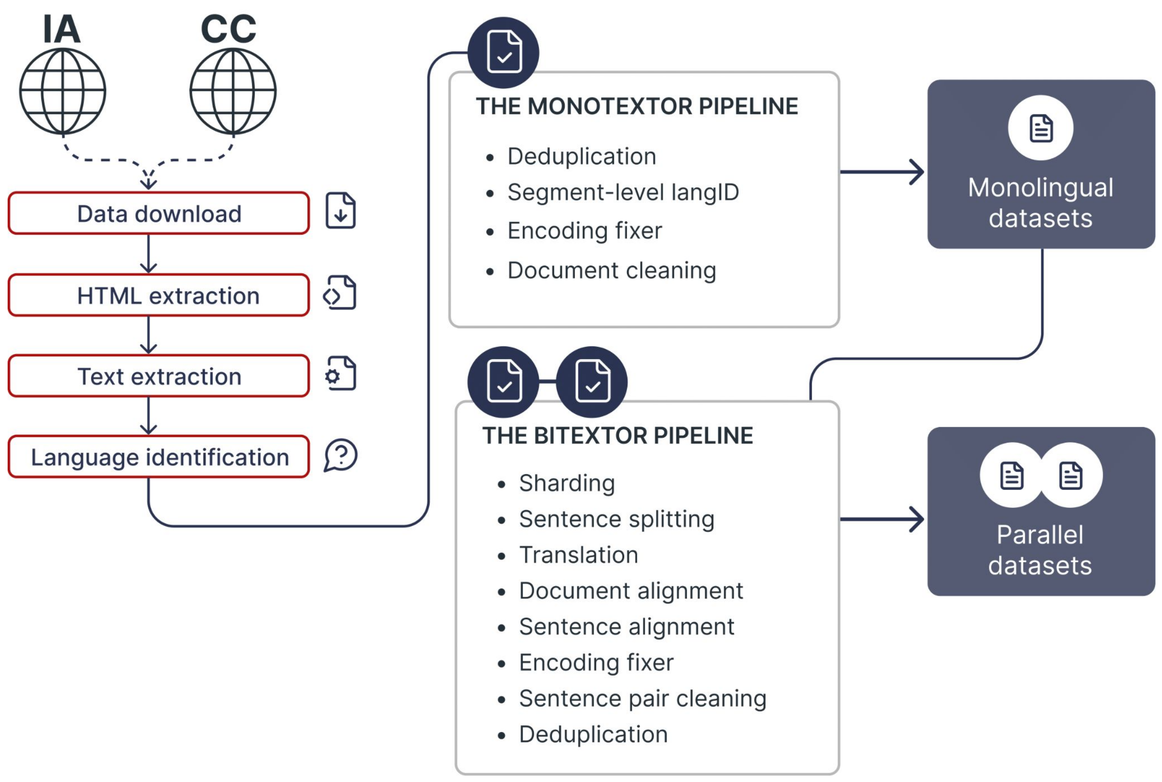
- **Monolingual:** 8 trillion tokens, 193 languages
- **Parallel:** 380M pairs, 50 languages X-En
- **DocHPLT v2:** 74M docs, 50 languages X-En
- **MultiHPLT v2:** 16.7B tokens, 1275 lang. pairs



Distribution of docs. in HPLT v2 monolingual cleaned corpora

Construction pipeline

Source: 4.5PB web crawl data (3.7 from IA, 0.8 from CC)
✨ HPLT is the **only** large-scale text collection extracted from IA.



Analysis and evaluation: demonstrating **quality** (much more in paper!)

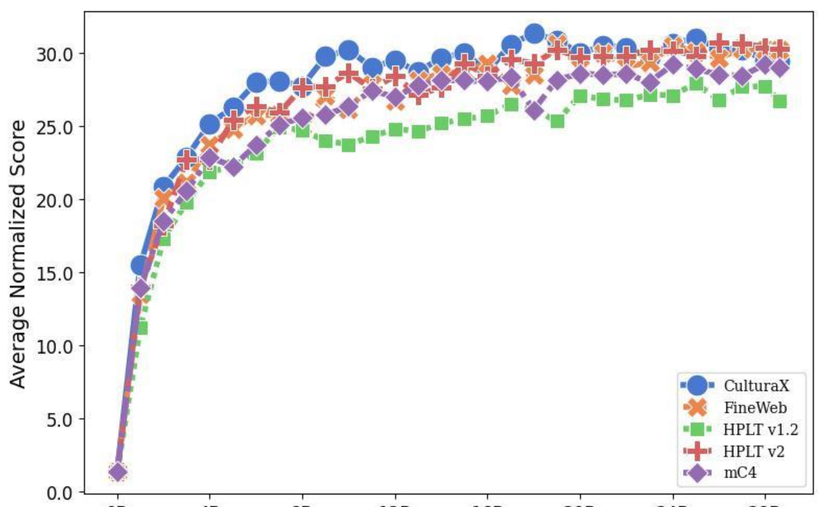
Manual data inspection

We inspected 50 randomly selected documents in 22 languages.

- <1% pornographic content
- 3% not target language
- 10% unnatural text

⇒ Pipeline works well and results in high-quality data

NLU tasks and large generative LMs



Performance for LLMs trained on Norwegian

⇒ HPLT v2: competitive and stable

Machine translation

	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.

⇒ Improvement over HPLT v1.2, complementary to existing data

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

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, Jaume Zaragoza-Bernabeu

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].