

Register-based data selection for multilingual LLMs

Tomasz Galica, Amanda Myntti, Erik Henriksson,
Fedor Vitiugin, Sampo Pyysalo & Veronika Laippala

TurkuNLP, University of Turku

{tomasz.galica, amanda.a.myntti, erik.henriksson,
fedor.vitiugin, spyysalo, mavela}@utu.fi

Abstract

The development of high-quality multilingual models depends on well-curated multilingual datasets, which are typically obtained through filtering large web-crawled corpora. Such filtering approaches must be effective across a wide range of languages to improve model quality proportionally across the entire linguistic space. In recent years, multiple Natural Language Processing (NLP) approaches for data processing and Large Language Model (LLM) training have been developed. However, the most significant progress has been made for rich-resource languages, such as English and Chinese, leaving medium and low-resource languages largely underexplored. In this work, we present data selection strategies based on Internet registers (genres) and Web Document Score (WDS). Utilizing HPLT 3.0 datasets, we study 15 European languages. We train and evaluate four small generative language models per language, and compare how specific registers and WDS affect model training in separate and combined approaches. Using five common evaluation tasks covering models’ linguistic, knowledge-based, and reasoning skills, we analyze the impact of the studied strategies on each language, their average performance, and show how combining register-based selection with WDS improves evaluation performance. Our results indicate that both registers and WDS can generalize across multiple languages, offering a new path towards multilingual data filtering.

1 Introduction

The performance of general-purpose multilingual and monolingual LLMs strongly depends on the quality, quantity, and composition of the pretraining data. Pretraining corpora must provide sufficient coverage and fluency of natural language, but also contain diverse linguistic characteristics and a broad range of human knowledge. Although multiple approaches for assessing data quality have been proposed, e.g., data filtering and deduplication (Tirumala et al., 2023; Penedo et al., 2024),

the evaluation of data diversity is also crucial to enable LLMs to learn different writing styles and cover a wide range of topics. Moreover, there is an interplay between data quality and diversity. As shown by Myntti et al. (2025), sampling data based on registers in English influences how the model performs. The best performance is achieved by combining individually high-performing registers.

Register, sometimes interchangeably known as genre in other frameworks, is a term widely adopted in corpus linguistics, specifically in the text linguistic approach to register variation (Biber, 2019). Register framework focuses on modeling language use and its variation across different situational contexts. The framework offers both an intuitive and theoretically grounded approach for categorizing web language use and sampling. In practice, register classes range from spoken conversations and opinionated think pieces to persuasive advertisements depending on the object of the study (Biber, 1988; Biber and Egbert, 2018; Biber and Conrad, 2019). Each register class also shows systematic differences in writing style, structure, and linguistic choices. In fact, the differences are so significant that register category can be considered as the most important predictor of linguistic variation (Biber, 2012). In our work, we use the same web register taxonomy as Laippala et al. (2023) and Henriksson et al. (2024), targeting registers on the entire unrestricted web. The approach has been adopted to label the HPLT 3.0 datasets that we use (Burchell et al., 2025).

This work builds on the methodology previously reported by Myntti et al. (2025). Their objective was to study how individual registers influence language models in English. The aim of our research is to propose a language-agnostic multilingual data filtering method, both for improved performance of language models and minimization of data loss. We study 15 languages other than English, showcasing the wide applicability of our work. We se-

lected a set of the largest HPLT 3.0 language subsets representing European national languages and languages relevant to the OpenEuroLLM project¹: Czech, Dutch, Finnish, French, German, Greek, Hungarian, Italian, Polish, Portuguese, Romanian, Spanish, Swedish, Turkish and Ukrainian.

We examine three data filtering methods: (1) register-based filtering, (2) Web document score (WDS) based filtering employing data upsampling and downsampling strategy, and (3) an approach mixing register and WDS methods. We utilize five model evaluation tasks and evaluate the impact of data selection strategies per language and present the averaged performance of the methods. Lastly, we examine the influence of the methods on the dataset size, which is crucial for low-resource languages.

2 Related work

Due to the conflict between the growing need for high-quality data for LLM training and the limits on its availability (Villalobos et al., 2024), dataset curation methods have become more necessary than ever before. Additionally, as shown by Ranathunga et al. (2024), significant variation in corpus quality can make random sampling an unreliable strategy for data selection, motivating quality-informed filtering. As a consequence, a multitude of studies into large-scale dataset composition optimization for LLM pretraining has emerged (Li et al., 2024; Penedo et al., 2024; Weber et al., 2024; Su et al., 2025; Nguyen et al., 2025). These studies often use simple statistical measures as heuristics or quality flags for selecting documents (Raffel et al., 2020; Rae et al., 2022), with Weber et al. (2024) using a mix of statistical flags and quality scores created by a small neural model. Furthermore, LLM-generated quality labels (Held et al., 2025; Henriksson et al., 2025) and synthetic training instances (Su et al., 2025) have been used in recent literature.

While large multilingual datasets have been released in recent years (e.g., Wenzek et al. (2020); Nguyen et al. (2024); Burchell et al. (2025)), most work on pretraining data curation has focused primarily on English. Recent research on multilingual data has begun to address this gap. Prominent examples of such studies include Wang et al. (2025), who use machine translation of the FineWebEdu (Penedo et al., 2024) dataset into nine languages;

Messmer et al. (2025) use embedding-based filtering methodology on 20 languages; Penedo et al. (2025), who construct the multilingual FineWeb2 dataset collection; and Ali et al. (2025), who use LLM-as-judge in a multilingual setting to train classifiers for selecting educational content from documents in 35 different languages.

Additionally, progress in multilingual studies has been constrained by the scarcity of high-quality multilingual benchmarks. This scarcity not only limits the reliability of the evaluation but also makes it difficult to perform cross-lingual comparisons. Many benchmarks are available for only a small subset of languages. Examples of recent work filling this gap include MultiBLiMP 1.0 (Jumelet et al., 2025), GlotEval (Luo et al., 2025), EuroEval² and the Multilingual FineTasks (Kydlíček et al., 2024).

Similar to LLM evaluation tasks, initial studies on automatic register identification were carried out for English first. Egbert et al. (2015) introduced the Corpus of Online REgisters (CORE), a dataset with a multilabel register classification scheme with a two-level hierarchy designed to cover the entirety of the English web. Laippala et al. (2023) developed the first deep learning-based automatic register identification model for the CORE scheme. The CORE scheme and automatic register identification have been extensively studied in a multilingual setting (Repo et al., 2021; Rönqvist et al., 2021), with Henriksson et al. (2024) releasing the Multilingual CORE collection encompassing register-annotated corpora for 16 diverse languages and an associated XLM-RoBERTa-based multilabel classifier. This classifier assigns labels based on the two-level hierarchy (with 9 main-level registers and 16 subregisters) and treats each register category as an independent binary prediction task. Additionally, Henriksson et al. (2024) show that register classification can be applied in a zero-shot fashion to languages not present in the classifier’s finetuning data. This classifier has been used to annotate the HPLT datasets with register labels (Burchell et al., 2025; Oepen et al., 2025), which form the basis of the register-based data selection in our work. Other work on multilingual register classification includes Kuzman et al. (2023) and Ljubešić and Kuzman (2024), who use the X-Genre scheme.

A central problem in LLM training is the lack of data transparency (Bommasani et al., 2025).

¹<https://openeurollm.eu/>

²<https://euroeval.com/>

One way to increase transparency in the pretraining phase is to provide metadata on the training datasets: for example, [Sharoff \(2020\)](#) uses a genre and a topic classifier to explain the contents of a corpus, and recent efforts such as the RedPajama dataset offer data accompanied by quality signals and metadata to enable analysis of dataset composition and filtering strategies ([Weber et al., 2024](#)). Register classification and other descriptive document metadata, such as the WDS, align with this standpoint.

3 Methods

3.1 Data

In our experiments, we use the HPLT 3.0 dataset collection ([Oepen et al., 2025](#)), which is a massive multilingual data repository comprising 198 languages sourced from the Internet Archive and Common Crawl. The majority of HPLT 3.0 datasets (104 languages) are annotated with probabilities of 25 main and sub register labels ([Henriksson et al., 2024](#)). Depending on the probability threshold, documents in HPLT 3.0 collection can have zero, one, or more main-level registers (known as a hybrid), and any associated subregister label. The full register-subregister scheme can be found in [Henriksson et al. \(2024\)](#).

Documents within HPLT datasets were also annotated by dataset creators with quality scores obtained with Web Doc Scorer (WDS)³. WDS is a language-agnostic framework designed to assess the quality of large-scale web corpora using surface-level and structural indicators. WDS assigns quality scores based on the well-formedness of the paragraph and the presence of linguistic data versus boilerplate or noise. Documents with excessive non-linguistic characters (e.g., numbers, emojis), high repetition rates, or a disproportionate amount of punctuation and URLs are penalized. The higher the WDS score, the higher the suitability of the document for model pretraining. WDS score ranges from 0 to 1, however, HPLT 3.0 maps this range to 0-10. HPLT 3.0 only includes documents with WDS score above or equal to 5.0. We utilize WDS scores as a second data filtering method.

3.2 Data processing

Our work was carried out using the following fifteen language datasets obtained from HPLT 3.0: Czech (ces), Dutch (nld), Finnish (fin), French (fra),

German (deu), Greek (ell), Hungarian (hun), Italian (ita), Spanish (spa), Polish (pol), Portuguese (por), Romanian (ron), Swedish (swe), Turkish (tur), and Ukrainian (ukr). To effectively study the filtering methods, we limited our study to cover languages with more than 50 billion tokens (BT) available for LLM training. In this work, we use the nine main-level registers and additionally two subregisters and one hybrid register, following [Myntti et al. \(2025\)](#) (see Table 1).

All datasets were processed in the same way. In the initial processing step, we classify documents according to the given probabilities with a classification threshold of 0.4. In the work of [Myntti et al. \(2025\)](#), the selection threshold is based on optimisation on the CORE dataset. While the optimal threshold for each language might be different, optimization requires register-annotated test sets, which do not exist for most of our languages. Thus, we use the same threshold of 0.4. Research by [Henriksson et al. \(2024\)](#) further strengthens our choice by noting that when the threshold is optimizable, it falls between 0.3 and 0.5.

For subregister classes News (ne) and Description (dpt), we retain the subregister label, and for the hybrid class Instructive-Informational (HI-IN), we assign this label if both values for How-to-Instructions (HI) and Informational description (IN) exceed the classification threshold. Otherwise, documents are left unlabeled and discarded.

Register label	Register description
HI*	How-to-instructions
ID*	Interactive discussion
IN*	Information description
IP*	Information persuasion
LY*	Lyrical
MT*	Machine translation
NA*	Narrative
OP*	Opinion
SP*	Spoken
HI-IN [†]	Instructive-Informational
ne [‡]	News
dtp [‡]	Description

Table 1: Registers used in our study and their classes:

* Main registers, [†] hybrid register, [‡] subregisters

We create three training subsets per language:

(1) **EX3 subset**, where we strictly excluded three registers that were reported to have the lowest performance for English data by [Myntti et al. \(2025\)](#): IP, MT, and LY. These cover advertisements, pre-GPT-style machine translations and lyrical texts, such as poems and song lyrics. By removing these registers, we examine whether the same register

³<https://github.com/pablop16n/web-docs-scorer>

categories have a similar impact across multiple languages. Unlike [Myntti et al. \(2025\)](#), we chose to exclude the lowest-performing registers rather than selecting the top-performing ones to retain more data. This is crucial in the multilingual setting where less data per language is available.

(2) **W subset**, for which we used WDS scores to filter the data. Instead of strictly removing the data with low scores, we used the continuous sampling function S (1) to obtain the sampling ratio. This function and its derivation is described in detail below this list.

(3) **RW subset**, for which we utilized both WDS and register labels to filter the data. We use the same S function as for the W subset. However, the final sampling ratio for the document is obtained by multiplying S by register multiplier R (described below the S strategy).

Next, we describe our up/downsampling strategy. Using a continuous function S , we downsampled (excluded) some of the documents with scores implying low-quality, and upsampled (repeated) documents with implied high quality. We developed this sampling strategy based on the following assumptions: (1) The maximal upsampling rate was fixed to 4 based on earlier research ([Taylor et al. \(2022\)](#), [Muennighoff et al. \(2023\)](#)) showing that data repetition rates up to 4 can be beneficial for model training, (2) The lowest downsampling value was set close to 0, (3) A third-order polynomial was adopted to create an S-shaped sampling profile; this ensures heaviest upsampling for high-quality data and significant penalization for low-quality data. Using a third-order polynomial, the function can easily fulfill these constraints with only four parameters. We select the values for the parameters based on the reasoning below.

To use the categorical registers in unison with WDS, we define register multiplier values R based on findings by [Myntti et al. \(2025\)](#): 0.1 for two registers with the lowest performance, IP and MT. The LY register, which also had low performance, is usually the smallest register across most languages, contributing less than 1% to the total dataset size. For this reason, we decided not to downsample it. The R was 1.5 for the top four performing registers for the English data: dtp, HI, HI-IN, and OP. The value of R was 1 for all other registers. To maintain linguistic variety, in both W and RW approaches we decided to not remove any register completely.

$$S = 0.0213x^3 - 0.4113x^2 + 2.9626x - 7.1492 \quad (1)$$

The final sampling function S , presented in Equation 1, indicates the percentage of documents excluded or repeated. Given maximal $R=1.5$ and the constraint of maximal total upsampling ratio being 4, the value of S was set to 2.67 on $WDS=10$. On medium-quality documents ($7 < WDS < 8$), the value of S is close to 1 (no up/downsampling), and on $WDS=5$, S is almost 0. The coefficients of S and register multipliers R were selected based on the given constraints and a limited number of initial experiments.

The baseline dataset and all the filtered subsets were subsequently tokenized with the GPT-OSS-120b tokenizer ([OpenAI, 2025](#)).

3.3 Model training

We used models with the Llama architecture [Touvron et al. \(2023\)](#) and model training was done with the NVIDIA Megatron-LM framework ([Shoeybi et al., 2019](#)). The models have 0.6 billion (0.6B) trainable parameters and were trained on up to 100 BT. In cases where dataset size was smaller than 100 BT, training was carried out for more than one epoch. The model architecture has 22 hidden layers with a hidden size of 1024, and 16 attention heads. We used a sequence length of 4096, a global batch size 1024, and Adam optimizer. Model hyperparameters were selected based on our initial experiments, and checkpoints were saved every 2000 steps. All computing was carried out on the CSC LUMI supercomputer. The training of a single model required, on average, 2250 GPU-hours, and the total computational cost of our study was 135000 GPU-hours (15 languages, 4 models per language).

We selected 0.6B parameter models as a compromise between computational efficiency and model capacity. Given the 15 languages and 4 models evaluated per language, larger models would substantially increase the computational cost, while smaller models could provide weaker and noisier signals of differences between data-selection strategies.

3.4 Evaluation

To evaluate the impact of our approach to data filtering across multiple languages, we tested our models with a broad range of evaluation tasks. However, only a limited number of evaluation tasks is available for multiple languages. To address this, we used two evaluation frameworks: LightEval ([Habib et al., 2023](#)) and LM-evaluation-harness

(Gao et al., 2024). Selected tasks are suitable for the 0.6B models we trained: they generally produce a monotonic signal above random baseline (see Section 4.2). The selected tasks are as follows:

- Belebele (Bandarkar et al., 2024) is a fully parallel multilingual multiple-choice reading comprehension task. The dataset contains 900 questions for each language. Questions were machine translated and human-annotated. This task was available via the LightEval framework with the evaluation dataset facebook/belebele⁴, and was available for all 15 languages that we studied.
- MultiBLiMP (Jumelet et al., 2025) is a benchmark of linguistic minimal pairs, testing subject-verb agreement for number, person, and gender. We used the LM-evaluation-harness implementation of this task based on the dataset jumelet/multiblimp⁵. All 15 languages were evaluated against this task and data.
- ARC: Challenge (Clark et al., 2018) is a task based on multiple-choice science exams for grades 3-9, testing multi-hop reasoning. Languages: ces, deu, ell, spa, fin, hun, ita, nld, pol, por, and swe were evaluated against LumiOpen/arc_challenge_mt⁶. Fra, ron, and ukr were evaluated against jon-tow/okapi_arc_challenge (Lai et al., 2023), and tur against malhajar/arc-tr⁷.
- HellaSwag (Zellers et al., 2019) is a sentence completion task for common sense reasoning evaluation. Deu, spa, fra, hun, ita, nld, por, ron, swe, and ukr were tested against jon-tow/okapi_hellaswag (Lai et al., 2023). LumiOpen/opengpt-x_hellaswag⁸ was used for ces, ell, fin, and pol, and malhajar/hellaswag_tr-v0.2 (Alhajar, 2024)⁹ for tur. All datasets are machine-translations from the English dataset.
- Global-MMLU (Singh et al., 2025) is a multilingual multiple-choice question task for testing different knowledge domains. Global-MMLU combines machine translation, professional, and

crowd-sourced annotation. It also addresses cultural sensitivity annotations, addressing western-based biases. For ces, deu, spa, fra, ita, nld, pol, por, ron, swe, tur, and ukr, we used the CohereForAI/Global-MMLU¹⁰ dataset. This Global-MMLU dataset was not available for ell, fin, and hun, and we substituted it with MMLU from LumiOpen/opengpt-x_mmlux (Hendrycks et al., 2021).

For Belebele, HellaSwag, ARC:Challenge, and Global-MMLU tasks, we used the cloze task formulation as implemented in the standard lighteval package. In the cloze (completion) task formulation, the model is presented with the task prompt and each of the answer options as completions, and the option with the highest likelihood is selected as the model’s answer. This contrasts with the multiple-choice formulation where all candidate answers are presented along with the prompt with labels (A, B, C...) and the model must select the correct label. All tasks were used as implemented in lighteval and LM-evaluation-harness packages.

The evaluation datasets used by us are well-known and publicly available, which introduces a risk of benchmark contamination and potentially inflated evaluation scores. However, all our models were trained on differently weighted versions of the same source data. Therefore, our evaluation focuses on relative performance between data-selection strategies rather than absolute model performance.

4 Results

In this section, we present results aggregated at two levels: average performance per language and per task. In addition, we present evaluation plots averaged for all languages and tasks for the baseline and the three data selection strategies. Baseline models are trained on unfiltered HPLT 3.0 data. We analyze the averaged model performance in relation to the proportion of data retained by each sampling method.

Our main experiments use 0.6B-parameter models. To assess whether the observed trends extend to larger models, we conducted an additional experiment with a 2B-parameter model trained on German. The results are presented in Appendix A figures 5, 6.

⁴<https://huggingface.co/datasets/facebook/belebele>

⁵<https://huggingface.co/datasets/jumelet/multiblimp>

⁶https://huggingface.co/datasets/LumiOpen/arc_challenge_mt

⁷<https://huggingface.co/datasets/malhajar/arc-tr>

⁸https://huggingface.co/datasets/LumiOpen/opengpt-x_hellaswag

⁹https://huggingface.co/datasets/malhajar/hellaswag_tr-v0.2

¹⁰<https://huggingface.co/datasets/CohereLabs/Global-MMLU>

4.1 Per language performance

The scores for each language were aggregated for base, EX3, W, and RW subsets, and the average score was calculated as the mean of five tasks. Figure 1 shows performance for each analyzed language. Plot markers show the calculated average score, and line plots show the exponential moving average with span=3.

We observe that for all languages, at least one of the data sampling strategies improves the average evaluation score over the baseline model, with the RW and EX3 strategies being the most successful. For 11 of the tested languages (ces, deu, fin, fra, hun, ita, nld, pol, spa, swe, and tur), the RW strategy consistently outperforms the baseline on all checkpoints. The W and EX3 strategies, while still successful, either achieve lower scores than RW or do not show such consistency during training.

For almost all models and languages, we observe a similar range of 0.4–0.46 averaged evaluation scores. This highlights the suitability of the selected tasks for comparison between different languages. The only exception is the Hungarian baseline and W models, which result in scores in the 0.32–0.35 range. In addition, for this language we observe a large gap between models: baseline and W models perform significantly worse than EX3 and RW. All four models were tested and confirmed to generate Hungarian. This unexpected behavior of baseline and W models might indicate the negative influence of IP and MT registers, which are either excluded or significantly downsampled in RW and EX3, but are present in baseline and W models.

4.2 Benchmark evaluation

To ensure that the used evaluation tasks produce a genuine informative signal, we studied the results of evaluation with quality metrics introduced by HuggingFaceFW Finetasks (Kydlíček et al., 2024):

- Monotonicity: measures whether the model is continuously learning as training progresses (see Table 2).
- Signal-to-noise ratio: describes whether the model produces a genuine signal or noise (see Table 3).
- Non-randomness: checks if the model score is above the random baseline (see Table 4).
- Ordering consistency: evaluates the model performance ranking consistency for all training checkpoints (see Table 5).

Language	base	EX3	W	RW
ces	0.9720	0.8811	0.9790	0.9510
deu	0.9720	0.9930	0.9161	0.9441
ell	0.6434	0.6224	0.9510	0.9021
fin	0.9580	0.9510	0.9231	0.8811
fra	0.9231	0.9510	0.8601	0.9720
hun	0.8182	0.9231	0.7692	0.9021
ita	0.8601	0.9790	0.9650	1.0000
nld	0.9720	0.9790	0.9371	0.9091
pol	0.9510	0.9161	0.9510	0.7203
por	0.9650	0.9441	0.9720	0.9860
ron	0.8671	0.9860	0.9790	0.9021
spa	0.9650	0.9441	0.9860	0.9720
swe	0.9510	0.9580	0.9021	0.9510
tur	0.8671	0.8671	0.7972	0.9650
ukr	0.8811	0.9650	0.8531	0.9161

Table 2: Monotonicity values for averaged evaluation tasks. Values closer to 1 indicate strictly increasing scores.

Language	base	EX3	W	RW
ces	101.69	65.65	103.29	106.72
deu	106.06	109.07	119.23	111.00
ell	163.65	89.02	169.29	183.81
fin	140.00	114.49	156.22	139.09
fra	85.04	96.10	83.37	100.31
hun	70.08	171.82	73.48	171.61
ita	71.40	102.59	128.65	107.84
nld	70.97	101.71	100.66	116.88
pol	147.36	126.38	119.92	111.84
por	95.95	103.23	97.30	120.90
ron	79.68	47.72	134.74	95.10
spa	87.94	89.46	106.62	94.33
swe	108.70	116.01	92.78	83.39
tur	59.38	76.32	90.30	71.10
ukr	107.19	101.65	108.90	102.82

Table 3: Signal-to-noise ratio for averaged evaluation tasks. HuggingFaceFW proposes a threshold of 20 as the minimum required to consider a model’s output to contain a meaningful signal.

Most of the models pass the quality tests. However, the Hungarian baseline and W models score very low on the non-randomness, and Greek, Swedish, Dutch, and Portuguese models have low ordering consistency. In particular Greek language score 0 in the consistency test. While RW model outperforms other three models for most checkpoints, there is no correlation between rankings of base, EX3, W and RW models across checkpoints. It suggest we cannot reliably predict models’ behaviors for Greek language beyond trained 100 BT. Despite this, in all cases we obtain high monotonicity scores and a good signal-to-noise ratio. Based on the analysis, we proceed with further analysis, not excluding any benchmark or language.

Language	base	EX3	W	RW
ces	0.1668	0.1650	0.1674	0.1706
deu	0.1617	0.1683	0.1648	0.1684
ell	0.1687	0.1724	0.1717	0.1741
fin	0.1373	0.1391	0.1386	0.1413
fra	0.1656	0.1698	0.1702	0.1771
hun	0.0688	0.1510	0.0599	0.1509
ita	0.1519	0.1584	0.1517	0.1631
nld	0.1588	0.1628	0.1578	0.1626
pol	0.1713	0.1748	0.1730	0.1767
por	0.1688	0.1723	0.1662	0.1697
ron	0.1600	0.1629	0.1608	0.1631
spa	0.1731	0.1776	0.1714	0.1816
swe	0.1604	0.1651	0.1641	0.1715
tur	0.1377	0.1429	0.1304	0.1491
ukr	0.1480	0.1540	0.1496	0.1540

Table 4: Non-randomness score for averaged evaluation tasks. The higher the score is above 0, the more distinct the signal is from the random baseline.

Language	Ordering consistency
ces	0.4444
deu	0.3333
ell	0.0000
fin	0.3889
fra	0.2222
hun	0.8333
ita	0.6667
nld	0.1111
pol	0.3333
por	0.1667
ron	0.4444
spa	0.3889
swe	0.0556
tur	0.8889
ukr	0.3333

Table 5: Ordering consistency per language, calculated for checkpoints above 50 BT. The higher the value, the better.

4.3 Per task and average performance

To summarize the performance of the data selection approaches, we aggregated and averaged the scores of all five evaluation tasks in Figure 2.

The results indicate that the RW strategy consistently improves performance across multiple task types. RW scores higher on both knowledge-based tasks, such as ARC:Challenge and Global-MMLU, as well as the linguistic reasoning MultiB-LiMP task. This suggests that the RW method not only helps select high-quality data but also preserves signals important for language structure and agreement. EX3, which excludes the lowest-performing registers, shows similar improvements in knowledge-focused and reading comprehension tasks. However, only the RW strategy shows a significant improvement over the baseline for the sentence-completion HellaSwag task. The

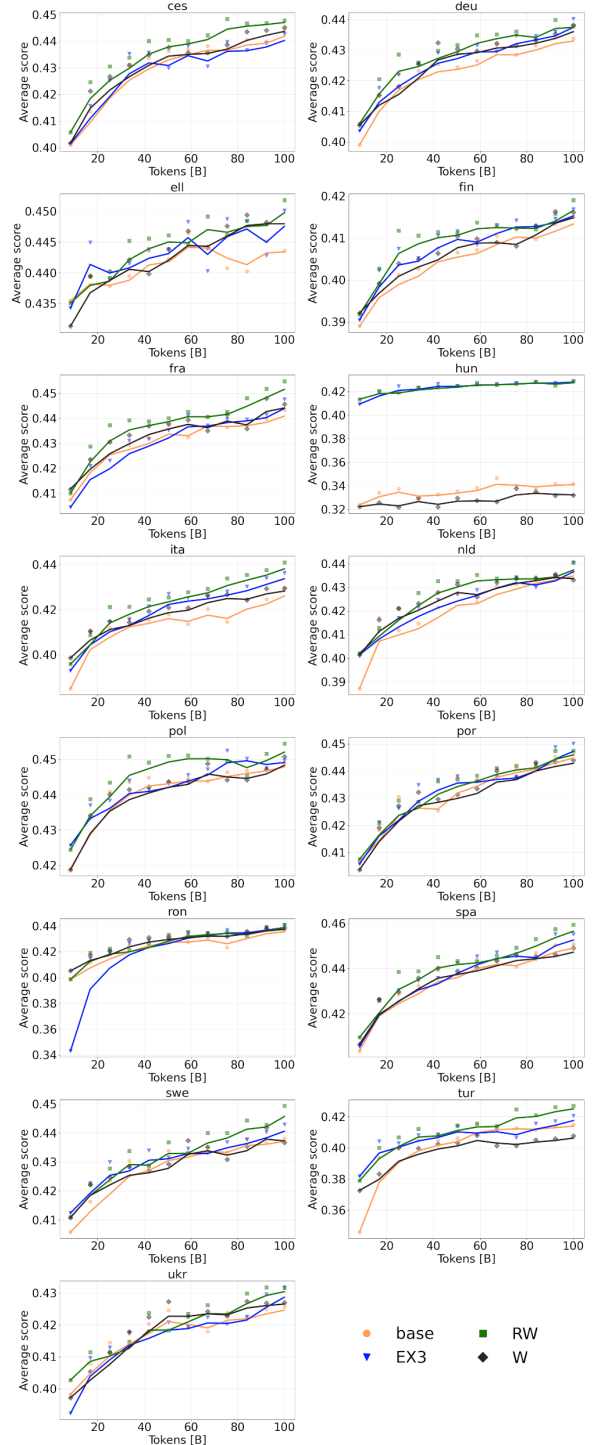


Figure 1: Performance of the monolingual models trained on baseline and sampled datasets per each tested language. Presented scores are average of 5 evaluation tasks.

W strategy improves performance primarily on knowledge-intensive tasks (ARC:Challenge and Global-MMLU) but has little benefit for linguistic or commonsense reasoning tasks. Due to Hungarian being an outlier in our previous tests, we compared the plots obtained for all 15 languages

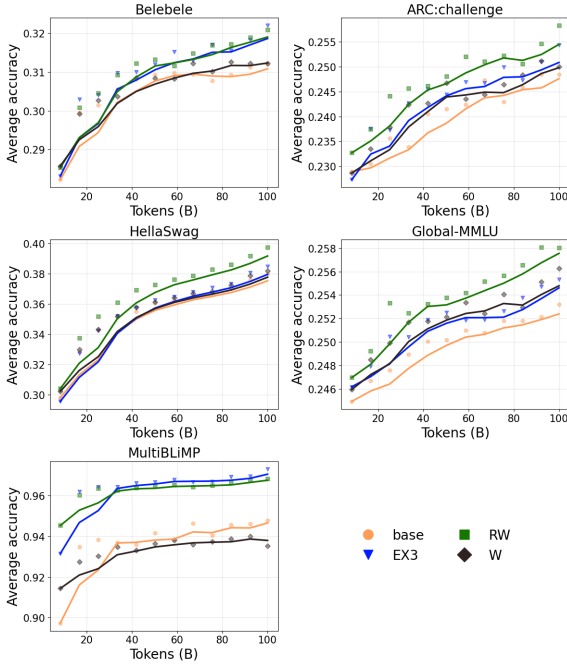


Figure 2: Evaluation plots for ARC:Challenge, Belebele, Global-MMLU, HellaSwag and MultiBLiMP tasks averaged over 15 languages.

to plots excluding Hungarian (see Appendix B Figure 7). Our interpretation does not change except for the MultiBLiMP task, where the EX3 strategy improves, and the RW approach achieves a similar score to the baseline.

Our findings suggest that out of the three data filtering methods, RW is the most generally robust strategy, as can be seen in the summary plot in Figure 3. The results presented in Figure 3 do not change significantly when Hungarian is excluded (see Appendix B Figure 8).

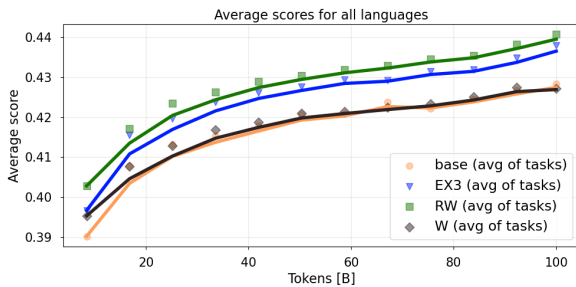


Figure 3: Averaged performance of baseline, EX3, W and RW data sampling strategies. Score is averaged over all languages and evaluation tasks.

4.4 Impact on dataset size

The focus of this work is on a multilingual approach to data selection. While data filtering is a

valid strategy for large languages, in the case of low-resource languages, the risk of data loss is a primary concern. Table 6 shows the sizes of individual datasets for all 15 languages. Figure 4 shows the sizes of subsets created by our data sampling approaches against the performance across the final 1-5 checkpoints. The pink dots refer to the performance of the last checkpoint, the bright red ones to the average score from the last 3 checkpoints, and dark red dots symbolize the average performance from the last 5 checkpoints. The lines connecting the dots show the minimal and maximal performance of the averages of the last 1-5 checkpoints. The RW sampling approach is clearly advantageous over other sampling approaches, balancing data loss and quality improvements.

While the EX3 approach performs significantly better than the baseline, it results in a substantial reduction of dataset size, removing 39% of the original data on average. The RW method results in subsets averaging 84% of the original size. However, the actual volume of removed documents in RW is higher than this reduction suggests. The final size is increased by the upsampling of the top performing registers and high WDS documents. Appendix B Table 7 presents the amount of documents removed for each language and method. The W approach has the least impact on dataset size, though it does not improve over baseline for our evaluation setting.

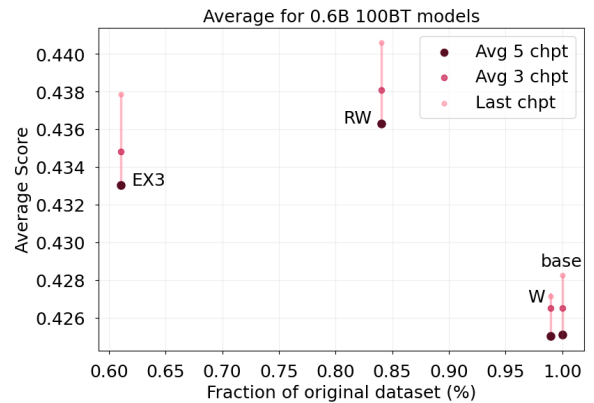


Figure 4: Performance vs fraction of baseline dataset size for baseline, EX3, W and RW data sampling strategies.

5 Conclusions

In this paper, we presented data selection strategies based on web register labels and WDS for multilingual datasets. We studied models trained on

	Size of datasets (BT)			
	Base	EX3	W	RW
ces	126.8	51.3	107.3	87.6
deu	498.7	284.3	504.18	397.9
ell	110.9	70.6	106.9	106.8
fin	67.8	37.7	62.3	50.2
fra	450.8	265.9	485.6	377.3
hun	101.4	60	105.4	88.2
ita	267.6	174.7	323.1	268.5
nld	122.4	78	122.8	102.2
pol	222.1	134	252.4	185.7
por	262.2	204.1	264.4	255.7
ron	101.4	63.6	98.2	89.3
spa	531.6	390	514.6	478.4
swe	105.6	55.5	83.9	66.9
tur	152.0	81.5	139.0	104.6
ukr	85.1	60.3	84.0	80.9
Avg. fraction	100%	61%	99%	84%

Table 6: Sizes of the unfiltered datasets and filtered subsets in billions of tokens (BT). The fractions compare the sizes of filtered models to unfiltered data.

100B tokens from the HPLT 3.0 dataset, with 0.6B trainable parameters. Our work was carried out for 15 languages, and confirms previous findings on English. Among the studied approaches, the mixed RW methodology outperformed EX3 and W strategies, and baseline models. While our continuous sampling approach might require additional optimization of sampling function coefficients, the consistent improvement of model quality across the languages highlights the generalizability of the method. Moreover, our research suggests that the approach presented by Myntti et al. (2025) and the dependencies discovered between the use of registers and model performance are transferable across the studied languages. We observed how the use of registers helped to filter out low-quality data for Hungarian, which dramatically improved evaluation scores. The proposed approach offers a good balance between data loss and quality improvement. However, to fully describe the impact of the proposed RW strategy, additional studies across a broader range of languages and model sizes above 0.6B trainable parameters are needed.

Limitations

The presented work offers a promising strategy for multilingual data filtering; however, despite generalizing to 15 languages, this study has a number of unexplored aspects.

Our work was carried out for the HPLT 3.0 dataset, and additional studies using different datasets might be needed to fully describe the transferability and generalizability of our approach.

HPLT 3.0 data is filtered (e.g., all documents are deduplicated, cleaned from HTML-artifacts and those with WDS below 5 were removed), and the impact of our approach on unfiltered data is not known. In addition, for all the studied languages, we assumed a uniform value for the register classification threshold. However, as shown by Henriksson et al. (2024), each language requires separate optimisation for this threshold. We also did not carry out register quality studies, assuming the transferability of the results described by Myntti et al. (2025). As a result, the ranking of top/bottom performing registers may be different depending on the language. Our work was based on European national languages and languages relevant to the OpenEuroLLM project, coming from three language families (Indo-European, Uralic, and Turkic) and using mainly Latin script. The only exceptions were Modern Greek and Ukrainian using the Greek and Cyrillic scripts, respectively. Further studies for non-European languages and non-Latin script languages should be carried out. This work focused on small language models with 0.6B trainable parameters due to resource feasibility. We conducted single study on 2B model. However, additional studies are needed to fully assess transferability of our results to larger models. Moreover, we carried out our study for monolingual models, and the effect of the RW approach on multilingual models is not yet known and is left for future research. While our analysis indicates that RW models outperform other approaches more in-depth studies with more varied evaluation tasks might be needed to fully describe the impact of the data selection on model capabilities. In addition, we did not carry out dataset decontamination against selected evaluation tasks, which means potential data contamination could have inflated our evaluation scores. However, as stated before, our analysis focuses on relative performance. Lastly, the benchmarks used are machine-translated and human-annotated. Using machine-translation as gold standard might not truthfully represent models’ capabilities to process target languages.

Acknowledgments

This project has received funding from OpenEuroLLM and LLMs4EU projects, under Digital Europe Programme grant agreements No 101195233 and 101198470, Finnish Doctoral Program Network in Artificial Intelligence (AI-DOC) under de-

cision number VN/3137/2024-OKM-6, and was supported by Centre of Excellence "Human Diversity through Contacts" (project 374223). Furthermore, this work was supported by the Research Council of Finland through FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union–NextGenerationEU instrument), and “Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora” (project362459). We gratefully acknowledge CSC– IT Center for Science Ltd. for providing computational resources. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

- Mohamad Alhajar. 2024. [Hellaswag turkish v0.2](#).
- Mehdi Ali, Manuel Brack, Max Lübbering, Elias Wendt, Abbas Goher Khan, Richard Rutmann, Alex Jude, Maurice Kraus, Alexander Arno Weber, Felix Stollenwerk, David Kaczér, Florian Mai, Lucie Flek, Rafet Sifa, Nicolas Flores-Herr, Joachim Koehler, Patrick Schramowski, Michael Fromm, and Kristian Kersting. 2025. [Judging quality across languages: A multilingual approach to pretraining data filtering with language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8870–8909, Suzhou, China. Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.
- Douglas Biber. 2012. [Register as a predictor of linguistic variation](#). *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.
- Douglas Biber. 2019. [Text-linguistic approaches to register variation](#). *Register Studies*, 1:42–75.
- Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, Cambridge.
- Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.
- Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2025. [The 2024 foundation model transparency index](#). *Transactions on Machine Learning Research*.
- 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, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. 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 answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Jesse Egbert, Douglas Biber, and Mark Davies. 2015. [Developing a user-based method of web register classification](#). *Journal of the Association for Information Science and Technology*, 66.
- 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, and 5 others. 2024. [The language model evaluation harness](#).
- Nathan Habib, Clémentine Fourier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. [Lighteval: A lightweight framework for llm evaluation](#).
- William Held, Bhargavi Paranjape, Punit Singh Koura, Mike Lewis, Frank Zhang, and Todor Mihaylov. 2025. [Optimizing pretraining data mixtures with llm-estimated utility](#). *Preprint*, arXiv:2501.11747.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 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.
- Erik Henriksson, Otto Tarkka, and Filip Ginter. 2025. [FinerWeb-10BT: Refining web data with LLM-based line-level filtering](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and*

- 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 258–268, Tallinn, Estonia. University of Tartu Library.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. [Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Preprint*, arXiv:2504.02768.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. [Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models](#). *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.
- Hynek Kydlíček, Guilherme Penedo, Clémentine Fourier, Nathan Habib, and Thomas Wolf. 2024. *Finetasks: Finding signal in a haystack of 200+ multilingual tasks*.
- Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. [Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback](#). *Preprint*, arXiv:2307.16039.
- Veronika Laippala, Samuel Rönqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. [Register identification from the unrestricted open web using the corpus of online registers of english](#). *Language Resources and Evaluation*, 57(3):1045–1079. Funding Information: Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital. Funding was provided by Academy of Finland (Grant No. 331297), Emil Aaltosen säätiö, National Science Foundation (Grant No. 1147581). Publisher Copyright: © 2022, The Author(s).
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. [Datacomp-LM: In search of the next generation of training sets for language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Nikola Ljubešić and Taja Kuzman. 2024. [CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282, Torino, Italia. ELRA and ICCL.
- Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Xu Huang, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofo, Fei Yuan, and Jörg Tiedemann. 2025. [GlotEval: A test suite for massively multilingual evaluation of large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 602–614, Suzhou, China. Association for Computational Linguistics.
- Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. 2025. [Enhancing multilingual llm pretraining with model-based data selection](#). *Preprint*, arXiv:2502.10361.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. *Scaling data-constrained language models*. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. 2025. [Register always matters: Analysis of LLM pretraining data through the lens of language variation](#). In *Second Conference on Language Modeling*.
- Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. 2025. [Recycling the web: A method to enhance pre-training data quality and quantity for language models](#). *Preprint*, arXiv:2506.04689.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [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.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. [Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). *arXiv preprint arXiv:2511.01066*.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language](#). In *Second Conference on Language Modeling*.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben al-lal, Anton Lozhkov, Margaret Mitchell, Colin Raffel,

- Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2022. [Scaling language models: Methods, analysis & insights from training gopher](#). *Preprint*, arXiv:2112.11446.
- 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. *J. Mach. Learn. Res.*, 21(1).
- Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. [Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian’s, Malta. Association for Computational Linguistics.
- Liina Repo, Valtteri Skantsi, Samuel Rönqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. [Beyond the english web: Zero-shot cross-lingual and lightweight monolingual classification of registers](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*.
- Samuel Rönqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. [Multilingual and zero-shot is closing in on monolingual web register classification](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Serge Sharoff. 2020. [Know thy corpus! robust methods for digital curation of web corpora](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2453–2460, Marseille, France. European Language Resources Association.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. [Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). *Preprint*, arXiv:2412.03304.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. [Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. [Galactica: A large language model for science](#). *Preprint*, arXiv:2211.09085.
- Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. [D4: Improving llm pretraining via document de-duplication and diversification](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53983–53995. Curran Associates, Inc.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. [Multilingual language model pretraining using machine-translated data](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28075–28095, Suzhou, China. Association for Computational Linguistics.
- Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish, and Ce Zhang. 2024. [Redpajama: an open dataset for training large language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, 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 *ACL (1)*, pages 4791–4800. Association for Computational Linguistics.

A Appendix

The main analysis presented in this work was carried out for models with 0.6 billion trainable parameters and for 100 BT training budget. However, to assess the effect of data filtering on larger models we carried out additional study for 2 billion trainable parameters German language (deu_Latn) model with 100BT training budget.

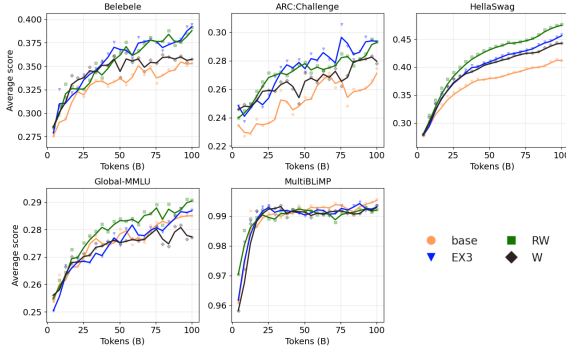


Figure 5: Evaluation plots for ARC:Challenge, Belebele, Global-MMLU, HellaSwag and MultiBLiMP tasks for deu 2B, 100BT model

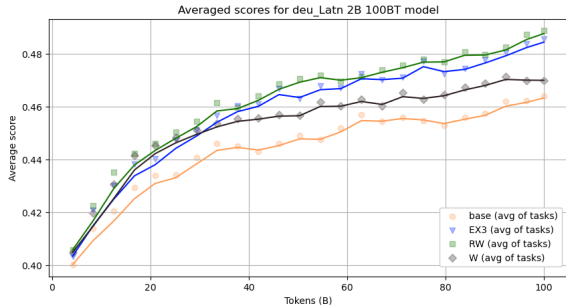


Figure 6: Averaged performance of baseline, EX3, W and RW data sampling strategies for 2B 100BT deu models.

Figures 5 show performance of the 2B deu_Latn models for the same set of tasks which was used

for 0.6B models: Belebele, ARC:challenge, HellaSwag, Global-MMLU, MultiBLiMP. The RW approach improves over other approaches for HellaSwag and Global-MMLU downstream tasks, and performs comparably well to EX3 for Belebele and ARC:challenge. The average score presented in figure and 6 shows a similar trend as seen in our main analysis in figure 3. The RW approach results in the highest average score, followed by EX3. However, in contrast to main study on 0.6B models, 2B deu_Latn W model has higher average performance than the baseline approach. The improvement is seen in Belebele, Arc:challenge and HellaSwag downstream tasks.

B Appendix

In this Appendix section, we present supplementary figures for the evaluation results when Hungarian language model is excluded from the average. Hungarian language model was found to be an outlier in non-randomness test (average score minimally above 0). In addition it displayed much larger difference in performance of baseline, W and EX3 and RW models.

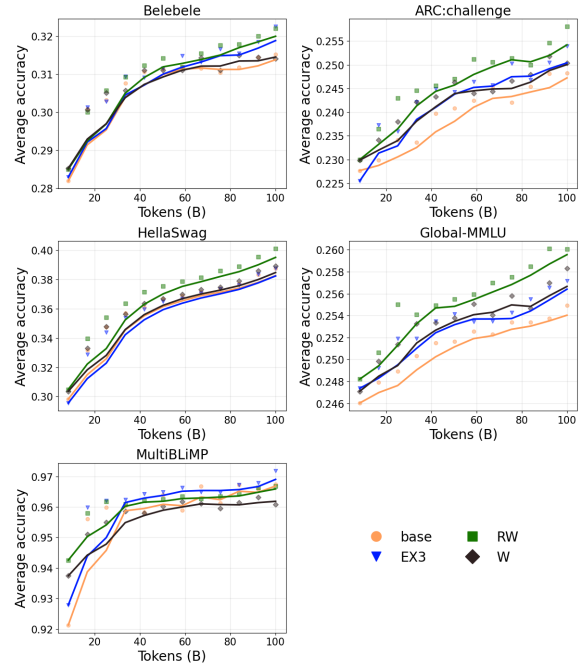


Figure 7: Evaluation plots for ARC:Challenge, Belebele, Global-MMLU, HellaSwag and MultiBLiMP tasks excluding Hungarian language

In the table 7, we present amount of documents removed by EX3, W and RW data filtering approaches. The final dataset sizes present in Table 6 is however effect of filtering out documents (re-

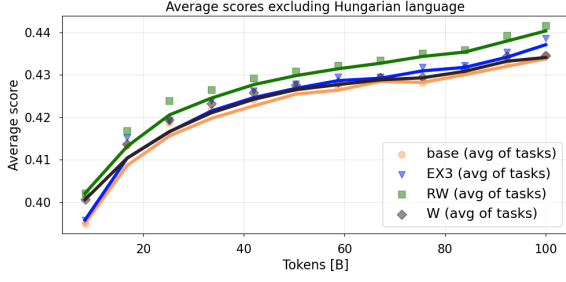


Figure 8: Averaged performance of baseline, EX3, W and RW data sampling strategies excluding Hungarian language.

moving them from dataset), and upsampling the documents with high quality signals.

	Amount of removed documents		
	EX3	W	RW
ces	51.9%	24.4%	48.4%
deu	50.4%	19.5%	54.3%
ell	28.3%	14.6%	71.5%
fin	46.3%	19.3%	58.1%
fra	50.7%	19.2%	54.6%
hun	44.9%	16.9%	60.6%
ita	42.8%	15.5%	62.1%
nld	48.7%	20.9	53.8%
pol	50.1%	19.7	52.5%
por	33.1%	17.8	66.2%
ron	40.0%	16.4	64.1%
spa	33.7%	20.9	63.7%
swe	46.3%	22.6	56.4%
tur	41.9%	19.6	61.2%
ukr	27.4%	16.2	68.8%
Avg. fraction	42.2%	18.9%	59.7%

Table 7: Amount of documents removed by EX3, W and RW filtering methods.

C Appendix

The distributions of WDS and registers for studied languages subsets of HPLT 3.0 are shown in tables 8 and 9 respectively. WDS distributions vary between languages, in most cases the highest amount of documents fall close to score 8.0. Figure 9 presents the WDS distribution averaged across studied languages. Similarly, we observe register variations between studied languages. Figure 10 shows the register distribution averaged for 15 studied languages. IP, ne and NA registers contain the largest shares of documents, while LY, SP and HI-IN containing the least amount of data. Additional HPLT 3.0 statistics can be found on the HPLT 3.0 website <https://analytics.hplt-project.org/viewer>. The differences in register distributions are associated with different classification threshold.

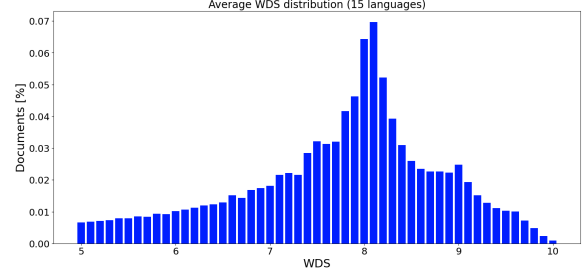


Figure 9: Averaged distribution of WDS across 15 studied languages

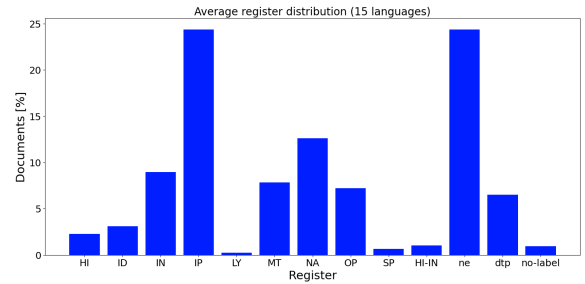


Figure 10: Averaged distribution of web registers across 15 studied languages

WDS	ces	deu	ell	fin	fra	hun	ita	nld	pol	por	ron	spa	swe	tur	ukr
5.0	0.93%	0.71%	0.42%	0.76%	0.74%	0.59%	0.50%	0.71%	0.68%	0.63%	0.56%	0.76%	0.81%	0.62%	0.54%
5.1	0.95%	0.73%	0.44%	0.77%	0.77%	0.59%	0.52%	0.74%	0.71%	0.64%	0.57%	0.78%	0.82%	0.64%	0.56%
5.2	0.99%	0.77%	0.48%	0.81%	0.78%	0.63%	0.55%	0.77%	0.74%	0.67%	0.59%	0.81%	0.87%	0.68%	0.58%
5.3	1.02%	0.79%	0.47%	0.80%	0.83%	0.64%	0.56%	0.78%	0.76%	0.69%	0.60%	0.83%	0.88%	0.68%	0.59%
5.4	1.11%	0.85%	0.54%	0.88%	0.84%	0.69%	0.62%	0.85%	0.83%	0.75%	0.66%	0.91%	0.97%	0.78%	0.63%
5.5	1.14%	0.85%	0.52%	0.92%	0.83%	0.70%	0.62%	0.83%	0.82%	0.75%	0.66%	0.88%	0.94%	0.74%	0.64%
5.6	1.20%	0.92%	0.60%	0.96%	0.88%	0.75%	0.68%	0.92%	0.90%	0.81%	0.72%	0.95%	1.04%	0.82%	0.69%
5.7	1.19%	0.91%	0.59%	0.88%	0.85%	0.75%	0.67%	0.90%	0.89%	0.80%	0.70%	0.93%	1.00%	0.81%	0.68%
5.8	1.31%	1.01%	0.70%	0.96%	0.94%	0.83%	0.76%	1.02%	1.00%	0.89%	0.79%	1.04%	1.13%	0.94%	0.76%
5.9	1.32%	1.00%	0.66%	0.96%	0.93%	0.83%	0.75%	1.00%	0.99%	0.89%	0.78%	1.03%	1.10%	0.92%	0.76%
6.0	1.43%	1.09%	0.74%	1.05%	1.00%	0.90%	0.82%	1.10%	1.08%	0.96%	0.85%	1.11%	1.20%	1.04%	0.83%
6.1	1.51%	1.13%	0.79%	1.06%	1.05%	0.96%	0.88%	1.15%	1.15%	1.02%	0.89%	1.17%	1.27%	1.13%	0.88%
6.2	1.57%	1.19%	0.83%	1.12%	1.12%	1.00%	0.92%	1.24%	1.20%	1.07%	0.93%	1.24%	1.33%	1.17%	0.92%
6.3	1.65%	1.25%	0.93%	1.17%	1.19%	1.08%	0.99%	1.30%	1.29%	1.12%	1.00%	1.29%	1.43%	1.27%	0.99%
6.4	1.71%	1.30%	0.91%	1.24%	1.24%	1.11%	1.03%	1.36%	1.33%	1.18%	1.05%	1.34%	1.46%	1.27%	1.03%
6.5	1.77%	1.34%	1.01%	1.27%	1.29%	1.17%	1.06%	1.41%	1.39%	1.23%	1.10%	1.39%	1.54%	1.39%	1.09%
6.6	2.00%	1.56%	1.17%	1.45%	1.51%	1.36%	1.24%	1.68%	1.65%	1.43%	1.31%	1.67%	1.82%	1.68%	1.26%
6.7	1.94%	1.48%	1.08%	1.41%	1.44%	1.32%	1.19%	1.56%	1.55%	1.37%	1.23%	1.54%	1.67%	1.54%	1.22%
6.8	2.16%	1.69%	1.34%	1.61%	1.67%	1.52%	1.37%	1.88%	1.80%	1.59%	1.43%	1.83%	2.08%	1.87%	1.41%
6.9	2.25%	1.79%	1.32%	1.70%	1.78%	1.62%	1.46%	1.94%	1.90%	1.67%	1.53%	1.84%	2.02%	1.85%	1.50%
7.0	2.26%	1.82%	1.49%	1.71%	1.83%	1.65%	1.54%	2.02%	1.91%	1.71%	1.58%	2.01%	2.22%	1.97%	1.55%
7.1	2.57%	2.09%	1.79%	2.06%	2.14%	1.95%	1.81%	2.47%	2.25%	2.01%	1.91%	2.37%	2.65%	2.39%	1.87%
7.2	2.65%	2.22%	1.71%	2.14%	2.28%	2.07%	1.88%	2.54%	2.33%	2.10%	2.05%	2.35%	2.56%	2.42%	2.00%
7.3	2.55%	2.14%	1.72%	2.06%	2.20%	2.06%	1.86%	2.44%	2.26%	2.05%	2.01%	2.30%	2.48%	2.29%	1.94%
7.4	2.95%	2.67%	2.68%	2.77%	2.74%	2.43%	2.42%	3.33%	2.85%	2.62%	2.64%	3.23%	3.65%	3.17%	2.52%
7.5	3.35%	3.14%	2.73%	3.10%	3.16%	2.87%	2.81%	3.86%	3.23%	2.96%	3.07%	3.53%	3.79%	3.61%	3.10%
7.6	3.21%	2.99%	2.82%	3.07%	3.16%	2.79%	2.78%	3.68%	3.05%	2.96%	3.07%	3.39%	3.52%	3.30%	3.13%
7.7	3.21%	3.09%	2.81%	3.13%	3.18%	2.87%	2.81%	3.77%	3.09%	3.07%	3.17%	3.49%	3.69%	3.46%	3.26%
7.8	3.62%	3.84%	4.48%	4.33%	3.79%	3.48%	3.65%	4.84%	3.73%	3.81%	3.91%	4.85%	5.36%	4.51%	4.34%
7.9	3.99%	4.30%	4.75%	4.93%	4.16%	4.04%	4.24%	5.32%	4.20%	4.30%	4.43%	5.10%	5.48%	4.90%	5.27%
8.0	4.69%	5.50%	10.07%	6.87%	4.91%	5.14%	5.69%	7.12%	5.15%	6.01%	5.71%	8.03%	8.16%	6.52%	6.92%
8.1	5.52%	6.41%	10.15%	7.58%	5.24%	6.10%	6.56%	7.90%	6.18%	6.78%	6.46%	7.37%	7.28%	7.15%	7.82%
8.2	4.45%	4.74%	7.00%	5.54%	4.22%	5.25%	5.15%	5.24%	4.76%	5.77%	5.33%	5.20%	4.49%	4.82%	6.38%
8.3	3.39%	3.54%	4.68%	4.03%	3.40%	4.15%	4.01%	3.85%	3.68%	4.62%	4.21%	4.00%	3.21%	3.40%	4.70%
8.4	2.84%	2.88%	3.31%	3.09%	2.96%	3.40%	3.32%	2.89%	3.01%	3.74%	3.43%	3.12%	2.39%	2.69%	3.46%
8.5	2.42%	2.51%	2.52%	2.57%	2.78%	2.92%	2.94%	2.25%	2.62%	3.12%	2.97%	2.48%	1.88%	2.18%	2.72%
8.6	2.19%	2.43%	2.09%	2.24%	2.75%	2.79%	2.83%	1.88%	2.49%	2.70%	2.79%	2.03%	1.57%	2.07%	2.39%
8.7	2.13%	2.51%	1.97%	2.13%	2.71%	2.79%	2.90%	1.72%	2.42%	2.50%	2.68%	1.79%	1.43%	2.08%	2.30%
8.8	2.05%	2.53%	1.96%	2.09%	2.68%	2.84%	3.09%	1.63%	2.44%	2.50%	2.73%	1.66%	1.35%	2.11%	2.27%
8.9	1.94%	2.49%	1.88%	2.00%	2.66%	2.81%	3.18%	1.56%	2.42%	2.46%	2.76%	1.57%	1.27%	2.11%	2.26%
9.0	1.93%	2.71%	2.51%	2.36%	2.97%	3.09%	3.55%	1.60%	2.67%	2.65%	3.13%	1.73%	1.34%	2.52%	2.36%
9.1	1.49%	2.33%	1.71%	1.57%	2.36%	2.46%	3.27%	1.10%	2.50%	1.87%	2.36%	1.05%	1.04%	2.07%	1.80%
9.2	1.17%	1.83%	1.40%	1.26%	1.98%	2.05%	2.35%	0.85%	1.87%	1.46%	1.97%	0.83%	0.81%	1.32%	1.55%
9.3	0.97%	1.51%	1.18%	1.09%	1.73%	1.78%	1.85%	0.68%	1.64%	1.25%	1.65%	0.70%	0.67%	1.04%	1.36%
9.4	0.84%	1.31%	1.05%	0.98%	1.52%	1.62%	1.48%	0.56%	1.36%	1.14%	1.44%	0.61%	0.58%	0.92%	1.25%
9.5	0.78%	1.18%	1.01%	0.95%	1.42%	1.59%	1.31%	0.51%	1.07%	1.12%	1.34%	0.58%	0.51%	0.86%	1.19%
9.6	0.72%	1.12%	1.09%	0.96%	1.40%	1.58%	1.31%	0.46%	0.96%	1.10%	1.29%	0.54%	0.45%	0.91%	1.14%
9.7	0.50%	0.82%	0.86%	0.72%	1.00%	1.16%	1.02%	0.32%	0.60%	0.77%	0.93%	0.37%	0.34%	0.64%	0.83%
9.8	0.30%	0.58%	0.62%	0.51%	0.62%	0.75%	0.73%	0.31%	0.36%	0.45%	0.60%	0.24%	0.26%	0.42%	0.52%
9.9	0.13%	0.32%	0.31%	0.27%	0.30%	0.35%	0.35%	0.11%	0.17%	0.19%	0.30%	0.11%	0.12%	0.21%	0.20%
10.0	0.05%	0.10%	0.12%	0.13%	0.12%	0.14%	0.10%	0.03%	0.08%	0.07%	0.11%	0.04%	0.06%	0.11%	0.06%

Table 8: Distribution of WDS scores in HPLT3 for 15 studied languages

Register	ces	deu	ell	fin	fra	hun	ita	nld
HI	2.79%	1.74%	1.90%	2.83%	2.24%	2.54%	2.23%	2.16%
ID	3.96%	4.81%	1.66%	3.79%	4.60%	3.20%	2.55%	2.78%
IN	6.14%	10.61%	4.95%	6.24%	11.04%	6.38%	9.69%	12.16%
IP	40.58%	26.16%	15.68%	26.14%	28.45%	25.95%	22.04%	25.88%
LY	0.14%	0.09%	0.25%	0.10%	0.26%	0.31%	0.19%	0.07%
MT	5.01%	15.83%	2.72%	9.68%	11.33%	8.57%	11.39%	9.12%
NA	11.54%	10.88%	15.85%	20.71%	10.35%	12.08%	11.97%	13.17%
OP	5.04%	7.16%	9.00%	4.99%	8.19%	6.13%	10.92%	8.15%
SP	0.50%	0.67%	0.68%	0.21%	1.17%	0.66%	0.99%	0.51%
HI-IN	0.89%	1.06%	0.42%	1.14%	1.16%	0.66%	0.92%	1.50%
ne	17.41%	13.48%	42.25%	18.69%	11.64%	27.47%	18.47%	13.32%
ntp	5.19%	6.61%	3.67%	4.60%	8.58%	5.21%	7.56%	10.43%
no-label	0.82%	0.90%	0.96%	0.87%	0.99%	0.84%	1.08%	0.75%

	pol	por	ron	spa	swe	tur	ukr
HI	1.96%	2.15%	2.33%	2.15%	2.12%	2.28%	2.65%
ID	6.26%	2.03%	2.07%	2.45%	2.46%	3.19%	0.59%
IN	10.90%	16.47%	4.84%	10.78%	7.18%	7.16%	9.64%
IP	32.33%	14.53%	24.19%	16.67%	26.32%	20.93%	19.44%
LY	0.18%	0.37%	0.34%	0.26%	0.08%	0.58%	0.18%
MT	6.22%	6.37%	4.00%	8.69%	8.54%	8.44%	1.35%
NA	11.21%	12.72%	10.57%	10.94%	25.10%	5.34%	6.53%
OP	6.99%	11.18%	6.71%	9.81%	5.16%	5.44%	3.14%
SP	0.84%	0.71%	0.50%	0.77%	0.27%	0.63%	0.39%
HI-IN	1.29%	1.14%	0.58%	1.23%	1.48%	0.84%	1.04%
ne	15.36%	22.60%	39.10%	24.60%	13.58%	37.78%	49.68%
ntp	5.51%	8.76%	4.01%	10.58%	6.80%	5.86%	4.54%
no-label	0.95%	0.97%	0.76%	1.07%	0.90%	1.53%	0.84

Table 9: Distribution of web registers in HPLT3 for 15 studied languages