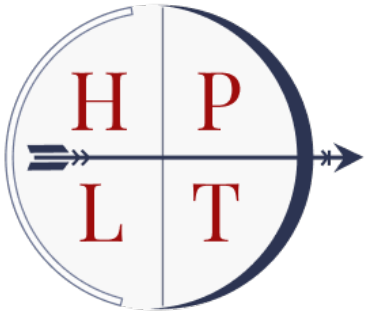
Co-funded by the European Union



HPLT: High Performance Language Technologies

Initial release of monolingual and parallel data sets

Deliverable number: 2.1

Version 1.1

Horizon Europe

UK Research and Innovation

Funded by the European Union's Horizon Europe search 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] programme

Project details

Project Acronym:	HPLT
Project Full Title:	HPLT: High Performance Language Technologies
Year of the Call:	2021
Type of Action:	HORIZON-IA (Innovation Action)
Grant Number:	101070350
Project URL:	https://hplt-project.org

Report details

Initial release of monolingual and parallel data sets	
Lead author:	Jörg Tiedemann (UH)
Contributing authors:	Nikolay Arefev (UiO) Andrey Kutuzov (UiO) Stephan Oepen (UiO) Jaume Zaragoza-Bernabeu (Prompsit) Mikko Aulamo (UH) Ona de Gibert Bonet (UH) Pavel Straňák (CUNI)
Internal reviewers:	Jan Hajič (CUNI) Gema Ramírez (Prompsit)
Deliverable number:	2.1
Dissemination level:	Public (PU)
Contractual Delivery Date:	Aug 31, 2023
Actual Delivery Date:	Aug 30, 2023
Number of pages:	32

Document history

Version	Date	Changes
1.0	Aug 31, 2023	Original Submission
1.1	Dec 05, 2023	Updated word count in table 4.2

Abstract

This report provides a description of deliverable D2.1 – the initial release of monolingual and parallel data sets coming from the HPLT project.

Contents

1 Executive summary

Deliverable D2.1 represents the initial release of raw monolingual and parallel texts (bitexts) acquired and compiled from web archives and crawls (1.7PB of data in total) by the HPLT project. The monolingual data collection covers 75 languages and a total of ≈ 21.7 trillion whitespace-separated word tokens. Bitexts focus on low resource languages and cover 14 language pairs and over 36 million aligned documents with roughly 6.1 billion tokens altogether. The releases are available with permissive licenses from our project website. The releases are complemented with open-source tools and pipelines used for processing huge web archive data packages. The data will be used in language and translation model training within HPLT and beyond and, therefore, represents an essential resource for the progress in the project and the wider research community. Note that the current release provides raw plain text data with some essential pre-processing but without fine-grained filtering, cleaning and de-duplication. Further data curation will be addressed in subsequent releases. Additionally, this document reports on the extension of OPUS with bitexts from various external resources. 981 resources have been added since the start of the project providing over 4 billion translation units. The import comes with a substantial reorganisation of the OPUS data hub and additional metadata such as overlap scores between different resources.

CHARLES UNIVERSITY

UNIVERSITY OF OSLO

UNIVERSITY OF EDINBURGH

UNIVERSITY OF TURKU

UNIVERSITY OF HELSINKI

PROMPSIT

CESNET

SIGMA2

2 Introduction

Deliverable D2.1 is a part of work package 2 (WP2) and concerns the release of monolingual and parallel bilingual text corpora extracted from large web crawls. This report complements the data release: it provides an overview of the data set, how they were produced and how we package and release them. Section 3 provides information about the process of data acquisition and our data management procedures. We describe the production of monolingual data in section 4 and the production and compilation of parallel data (the so-called 'bitexts') in section 5. Finally, section 6 provides details about packaging and release information.

D2.1 is the first of two data releases planned in WP2 in the project and constitutes the initial release of newly created data sets from unrestricted web crawls and web archive data. The deliverable is mostly connected with all four tasks in the WP including work on storage and compute infrastructure (Task 2.1), the acquisition of large quantities of archived web data on the scale of petabytes (Task 2.2), the extraction of monolingual plain text data with a focus on lesser-resourced languages (Task 2.3) and the extraction, alignment and compilation of parallel data in a consistent format (Task 2.4). It also reflects part of the work done in WP3 related to getting cleaning tools ready (Task 3.1) and applying them to produce data releases (Task 3.2). For this release, only some basic cleaning steps have been applied, and, hence, the released data contains a good amount of noise still to be filtered due to the nature of the original data (unrestricted internet archives and web crawls). Monolingual data and bitexts are not de-duplicated either at this stage. Providing data at this stage enables application-specific data selection and pre-processing routines, and does not dictate specific procedures on subsequent pipelines. Further data curation will be applied in subsequent data releases leading to cleaner collections that can directly be used for model training but any user should be aware that the data from this initial release require further refinements to be effective in language and translation modeling.

The report includes also information about data collection from other resources. In particular, we describe the extensions of OPUS with additional public parallel data sets. Those bitexts have been converted and imported into our collection continuing our efforts to produce a unified data hub for wide-coverage parallel data. The import procedures are available through GitHub and we can report 981 new resources providing more than 4 billion translation units. In connection with this extension, we also improved the structure of the data collection and added necessary metadata in a consistent format. More details will be given in Section 5.6.

Below, we provide information about the procedures, statistics of the data sets and links to the resources that we release. Furthermore, we also publish the pipelines, scripts and tools that enabled the extraction and compilation of the data sets.

The partners involved in this deliverable and the underlying tasks are:

- **CESNET** and **SIGMA2** for task 2.1 addressing storage and compute infrastructures
- **University of Oslo** for task 2.2 and task 2.3 with a focus on the acquisition of web archive data and the creation of monolingual plain text corpora
- **University of Edinburgh** for tasks 2.2, 2.3 and 2.4 focusing on monolingual and parallel text extraction

- **University of Helsinki** for task 2.4 with the extraction of bitexts, data ingestion into the public data collections and for coordination in WP2
- **Prompsit** for tasks 3.1 and 3.2 with responsibilities for the data pipelines, tool setups and also for the actual processing of data
- **CUNI** for providing the LINDAT/CLARIAH-CZ repository for permanent metadata record and persistent ID for the data, and for setting the license terms.

Data releases are available through links from our project website[1] and OPUS.[2] Software and tools are released through our workspace at GitHub.[3]

[1] https://hplt-project.org/datasets/
[2] https://github.com/Helsinki-NLP/OPUS
[3] https://github.com/hplt-project

3 Data Acquisition and Management

3.1 Source data

3.1.1 Data sources and storage facilities

Data acquisition in HPLT relies on two main sources of web crawls: the Internet Archive[1] and Common Crawl[2]. The computing and storage resources of NIRD[3] in Norway and CESNET[4] in Czech Republic were used to download and pre-process web crawls from these two sources.

For the current data release we have downloaded and stored on NIRD two large web crawls from the Internet Archive (IA) named WIDE15 and WIDE17 along with the CC-MAIN-2022-40 (CC40) crawl from Common Crawl. These crawls occupy a total of 1082 TB on NIRD. The download speed varied between 9 and 33 TB/day. On CESNET, we downloaded a third crawl from Internet Archive, WIDE16, which accounts for 768 TB. The download speed was 17.5TB/day.

Web crawls from both Internet Archive and Common Crawl are available in the WARC (Web Archive) format.[5] A WARC file stores communication messages between a web server and a crawler. It is composed of WARC records of several types, the main types are requests sent by a crawler to a server and responses it received from the server.

3.1.2 Characteristics of web crawls employed

Crawl	CC40	IA WIDE15	IA WIDE16	IA WIDE17
WARC size, TB	83	358	768	641
#WARC items	-	38782	83466	69386
#WARC files	80000	361431	754143	662381
≈ time to download, days	2.5	40	27	28
Download speed, TB/day	33.20	8.95	17.5	22.89
Download threads	20	128	256	2000

Table 3.1: Web crawls behind the initial data release in HPLT.

Table 3.1 provides an overview of characteristics of individual crawls. Common Crawl provides a list of URLs from which WARC files can be directly downloaded. Internet Archive provides a hierarchical structure where WARC files are grouped into *items*, and a crawl consists of a number of items. For a crawl from Internet Archive we store each item in a separate directory containing all corresponding WARC files, while for a crawl from Common Crawl all WARC files are stored in a single directory.

The download speeds and times specified in the table vary significantly from crawl to crawl due to the development of the downloading scripts and tuning of their parameters, mainly the number of

[1] https://archive.org/
[2] https://commoncrawl.org/
[3] https://www.sigma2.no/data-storage
[4] https://www.cesnet.cz/
[5] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1-annotated/

downloading threads. We expect it to be on the higher end for the future downloads.

3.2 Structure of the stored data

In the directory structure developed for storing data, the downloaded WARC files are grouped into directories by their corresponding crawls and data sources, and stored under the *one/warc* directory (see table 3.2), where 'one' corresponds to the first data release. Texts extracted and processed as described in sections 4.1 and 4.4 are stored under *one/text* and *one/monotexted* correspondingly. The code used to produce the first data release is in *one/code*.

Directory	Description
one	the first data release
one/warc	WARC files for the first data release
one/warc/ia	crawls from the Internet Archive
one/warc/ia/wide00015	
one/warc/ia/wide00016	
one/warc/ia/wide00017	
one/warc/cc	crawls from the Common Crawl
one/warc/cc/cc40	
one/text	extracted texts for the first data release
one/text/ia	texts extracted from the Internet Archive
one/text/ia/wide00015	
one/text/ia/wide00016	
one/text/ia/wide00017	
one/text/cc	texts extracted from the Common Crawl
one/text/cc/cc40	
one/code	code employed for preparing the first data release
one/monotexted	final data to be included in the first data release
one/monotexted/cc40	
one/monotexted/wide15	
one/monotexted/wide16	
one/monotexted/wide17	

Table 3.2: Directory structure developed for storing data related to the first data release

3.3 Downloading scripts

Scripts for downloading crawls from Internet Archive and Common Crawl were developed and published in the HPLT git repository.[6] Important features of those scripts include downloading data in multiple threads and automatically reattempting to get the files that failed to download correctly after some time. These features are vital for downloading large file collections such as web crawls. They will facilitate downloading new crawls for future releases.

[6] https://github.com/hplt-project/ia-download

4 Monolingual Data

This chapter starts with data processing procedures which are common for both monolingual and bilingual datasets. Beginning from the section 4.4, it focuses on monolingual data.

4.1 Extracting raw texts from WARC files

4.1.1 The `warc2text` tool

The downloaded crawls were processed by the `warc2text` tool[1] from the Bitextor pipeline in order to extract raw texts for each of the supported languages. A WARC file stores interactions between a web crawler and web servers storing web sites to be crawled, consisting of requests generated by the crawler and responses to these requests obtained from servers. Among all stored responses, `warc2text` finds documents containing text in some natural language and does fast preliminary filtering of undesirable documents based on their URL or HTML tags. More thorough filtering happens at the next stages. From the remaining documents, it extracts raw unformatted text and detects its language. Paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) are replaced by newlines in this raw text. The output of `warc2text` consists of compressed base64-encoded raw texts along with the URLs of the original web pages these texts originate from. This data is grouped into directories by language, which is detected using the CLD2 language classifier[2].

4.1.2 Text extraction with `warc2text`

Since `warc2text` is a part of Bitextor and lacked a release versioning policy at the time of running this part of our workflow, we report the commits of `warc2text` that were used, that is, commit '1b2e248' on NIRD and commit 'eac887e' on CESNET.

Table 4.1 presents general information about texts extracted from each crawl. The texts themselves are stored in `text.gz` files accompanied by `url.gz` files containing the original URLs they were crawled from. These files are grouped into directories corresponding to the detected languages. Each `warc2text` task takes a list of WARC files and generates as many directories as languages were detected. On the upper level, we run multiple `warc2text` tasks in parallel and have a separate directory of this structure for each task. At the next processing step this data is restructured to reduce the number of files and balance data across batches.

For the first data release, 77 languages were selected and all documents determined to be in other languages were filtered out. See Table 4.2 for the list of selected languages. Since the most popular languages were selected, this resulted only in minor reduction of total uncompressed text size. The only exception is WIDE16 which was processed by a version of `warc2text` that puts documents in undetected language to the `UNK` directory instead of skipping them, thus, filtering out this category resulted in about 20% reduction in size.

[1] https://github.com/bitextor/warc2text
[2] https://github.com/CLD2Owners/cld2

Crawl	CC40	IA WIDE15	IA WIDE16	IA WIDE17
# files after `warc2text`	384360	1490152	1955584	2403058
compressed text size, TB	8.4	19	42	18
uncompressed text size, TB	18.04	38.15	130.82	43.65
uncompressed text size for 77 languages, TB	18.00	38.12	103.44	43.62
# text.gz files	127853	495512	977792	798811
`warc2text` time, hours	23	38	500	48
`warc2text` threads	245	245	60	245

Table 4.1: Raw texts extracted from crawls by the `warc2text` tool

4.1.3 Per-language statistics

Table 4.2 shows the volumes of texts extracted by `warc2text` for each language. The number of segments (lines), words and bytes are as reported by the Unix `wc(1)` tool, see its documentation for definitions of a line and a word. Each WARC record is counted as a single document, different records having the same URL or text are counted as different documents. The volume of texts significantly ranges from 2.2 GB for text classified by CLD2 as Esperanto to 77.5 TB for English, while the number of documents has the minimum of 314K for Pashto and the maximum of 12.8B for English. We foresee a high percentage of documents mis-classified by CLD2 due to the huge amount of noisy data that it receives at this stage.

Language	Code	# Segments	# Words	# Bytes	# Documents
Esperanto	eo	5.54e+07	2.91e+08	2.08e+09	4.01e+05
Pashto	ps	6.03e+07	2.69e+08	2.82e+09	3.14e+05
Tatar	tt	6.40e+07	1.67e+08	2.98e+09	3.70e+05
Kyrgyz	ky	5.15e+07	1.81e+08	3.49e+09	3.45e+05
Somali	so	8.02e+07	5.40e+08	3.65e+09	6.78e+05
Welsh	cy	1.33e+08	6.20e+08	4.29e+09	7.35e+05
Irish	ga	3.56e+08	1.43e+09	1.00e+10	2.73e+06
Maltese	mt	3.61e+08	1.39e+09	1.01e+10	9.31e+05
Basque	eu	3.38e+08	1.51e+09	1.15e+10	2.29e+06
Norwegian Nynorsk	nn	3.99e+08	1.62e+09	1.17e+10	1.86e+06
Uzbek	uz	2.18e+08	9.25e+08	1.21e+10	1.37e+06
Gujarati	gu	2.28e+08	7.41e+08	1.33e+10	9.28e+05
Swahili	sw	3.72e+08	2.07e+09	1.37e+10	2.20e+06
Punjabi	pa	3.28e+08	1.14e+09	1.61e+10	2.41e+06
Nepali	ne	2.63e+08	9.32e+08	2.44e+10	2.12e+06
Icelandic	is	9.11e+08	3.41e+09	2.62e+10	3.87e+06
Marathi	mr	3.56e+08	1.31e+09	2.71e+10	1.64e+06
Mongolian	mn	4.89e+08	1.59e+09	2.93e+10	2.51e+06
Kannada	kn	7.59e+08	1.79e+09	2.96e+10	1.94e+06
Sinhalese	si	3.35e+08	1.57e+09	3.01e+10	1.37e+06
Kazakh	kk	5.90e+08	1.90e+09	3.38e+10	3.52e+06
Galician	gl	1.23e+09	4.96e+09	3.40e+10	4.62e+06
Macedonian	mk	8.23e+08	2.32e+09	3.48e+10	3.43e+06
Telugu	te	5.37e+08	1.94e+09	3.60e+10	3.58e+06
Malayalam	ml	5.26e+08	1.94e+09	4.33e+10	2.19e+06
Afrikaans	af	1.43e+09	6.08e+09	4.54e+10	4.39e+06
Tagalog	tl	1.77e+09	7.04e+09	4.75e+10	4.99e+06
Armenian	hy	1.16e+09	3.23e+09	4.79e+10	4.00e+06

Urdu	ur	1.46e+09	5.54e+09	5.06e+10	6.13e+06
Burmese	my	1.05e+09	3.62e+09	6.14e+10	2.47e+06
Belarusian	be	9.26e+08	4.29e+09	6.82e+10	3.93e+06
Albanian	sq	2.25e+09	1.15e+10	8.12e+10	9.20e+06
Azerbaijani	az	2.86e+09	1.06e+10	9.17e+10	1.05e+07
Latin	la	2.55e+09	1.43e+10	1.05e+11	2.06e+07
Georgian	ka	2.16e+09	6.37e+09	1.05e+11	6.52e+06
Tamil	ta	1.53e+09	6.04e+09	1.58e+11	5.46e+06
Slovenian	sl	5.60e+09	2.28e+10	1.61e+11	1.92e+07
Catalan	ca	5.28e+09	2.42e+10	1.62e+11	2.43e+07
Estonian	et	6.30e+09	2.16e+10	1.71e+11	2.13e+07
Croatian	hr	8.05e+09	3.14e+10	2.21e+11	2.80e+07
Latvian	lv	7.04e+09	2.73e+10	2.22e+11	2.13e+07
Lithuanian	lt	8.48e+09	3.29e+10	2.54e+11	3.27e+07
Serbian	sr	1.02e+10	3.51e+10	2.93e+11	3.22e+07
Malay	ms	1.16e+10	5.46e+10	3.53e+11	2.84e+07
Norwegian Bokmål	no	1.52e+10	5.50e+10	3.81e+11	6.19e+07
Hindi	hi	8.88e+09	3.47e+10	4.89e+11	3.39e+07
Hebrew	iw	1.50e+10	4.62e+10	6.04e+11	4.68e+07
Ukrainian	uk	1.27e+10	4.08e+10	6.59e+11	4.73e+07
Swedish	sv	2.36e+10	9.46e+10	6.65e+11	9.80e+07
Finnish	fi	2.43e+10	7.98e+10	6.75e+11	8.93e+07
Slovak	sk	2.89e+10	9.44e+10	6.95e+11	9.08e+07
Bengali	bn	1.49e+10	4.30e+10	7.19e+11	2.34e+07
Bulgarian	bg	1.58e+10	5.35e+10	7.44e+11	5.89e+07
Danish	da	4.16e+10	1.39e+11	9.50e+11	6.91e+08
Romanian	ro	3.29e+10	1.44e+11	1.12e+12	1.05e+08
Hungarian	hu	3.76e+10	1.37e+11	1.13e+12	1.39e+08
Czech	cs	5.23e+10	1.92e+11	1.38e+12	1.85e+08
Thai	th	3.22e+10	7.29e+10	1.39e+12	9.56e+07
Indonesian	id	2.95e+10	2.07e+11	1.45e+12	1.28e+08
Korean	ko	5.56e+10	1.36e+11	1.57e+12	2.49e+08
Vietnamese	vi	4.53e+10	2.84e+11	1.74e+12	1.75e+08
Turkish	tr	5.03e+10	2.37e+11	1.83e+12	2.17e+08
Dutch	nl	6.20e+10	2.49e+11	1.83e+12	2.36e+08
Arabic	ar	4.01e+10	1.81e+11	2.20e+12	1.99e+08
Greek	el	5.43e+10	1.78e+11	2.57e+12	1.36e+08
Polish	pl	8.71e+10	3.65e+11	2.66e+12	3.50e+08
Italian	it	8.82e+10	4.02e+11	2.80e+12	3.43e+08
Persian	fa	6.31e+10	2.61e+11	2.93e+12	1.91e+08
Chinese traditional	zh-Hant	7.42e+10	1.25e+11	3.14e+12	4.74e+08
Portuguese	pt	1.26e+11	5.96e+11	3.92e+12	4.53e+08
French	fr	1.73e+11	7.84e+11	5.27e+12	6.66e+08
Spanish	es	1.62e+11	8.51e+11	5.53e+12	6.81e+08
Japanese	ja	1.75e+11	2.59e+11	6.10e+12	6.85e+08
German	de	2.23e+11	8.98e+11	6.87e+12	1.03e+09
Russian	ru	4.34e+11	1.42e+12	2.18e+13	1.54e+09
Chinese simplified	zh	8.14e+11	1.44e+12	3.75e+13	6.50e+09
English	en	2.20e+12	1.02e+13	6.84e+13	1.28e+10
Total		5.40e+12	2.07e+13	1.95e+14	2.91e+10

Table 4.2: Raw texts extracted with `warc2text` per language using CLD2: the number of segments (new line symbols), words (as defined by `wc(1)`), bytes and documents. Ordered by size in bytes.

Figures 4.1 and 4.2 describe the proportions of data coming from each crawl and language. While for most languages the majority of texts come from the largest crawl, WIDE16, for Chinese the main source is WIDE17 and Esperanto, Basque, Nepali come primarily from CC40 even though it is five

times smaller than WIDE16. Thus, a combination of different crawls including small ones seem to be beneficial for covering different languages reasonably well.

4.2 Code for text extraction and statistics calculation

We have developed code that runs `warc2text` in many parallel threads, which is essential for processing large collections of WARC files as the ones that we have. In addition, code for calculating and plotting different statistics for those crawls including statistics presented in Section 4.1 was developed to facilitate further analysis. The code is publicly available.[3]

[3] https://github.com/hplt-project/warc2text-runner

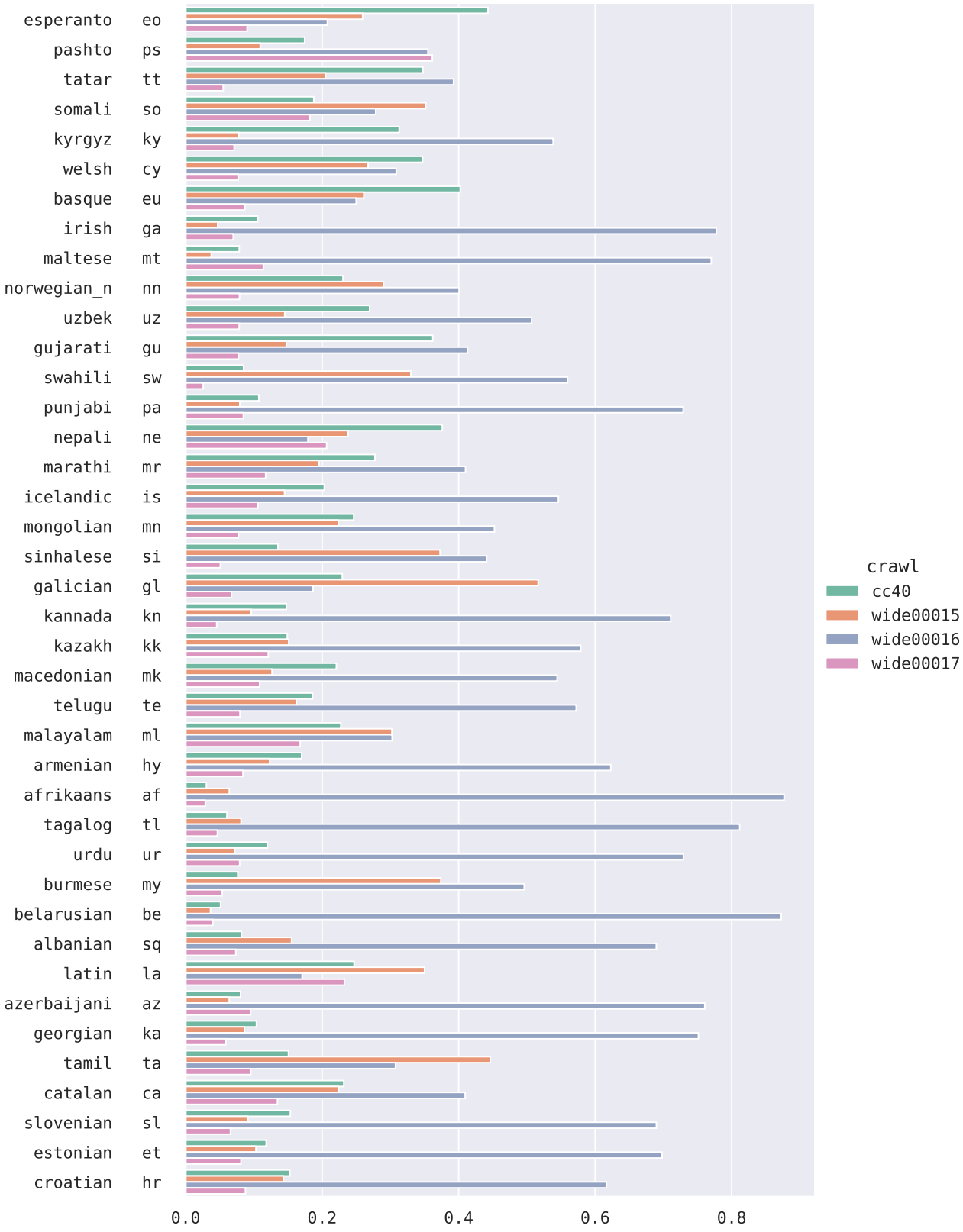


Figure 4.1: Proportions of text volume in bytes coming from each crawl, part 1.

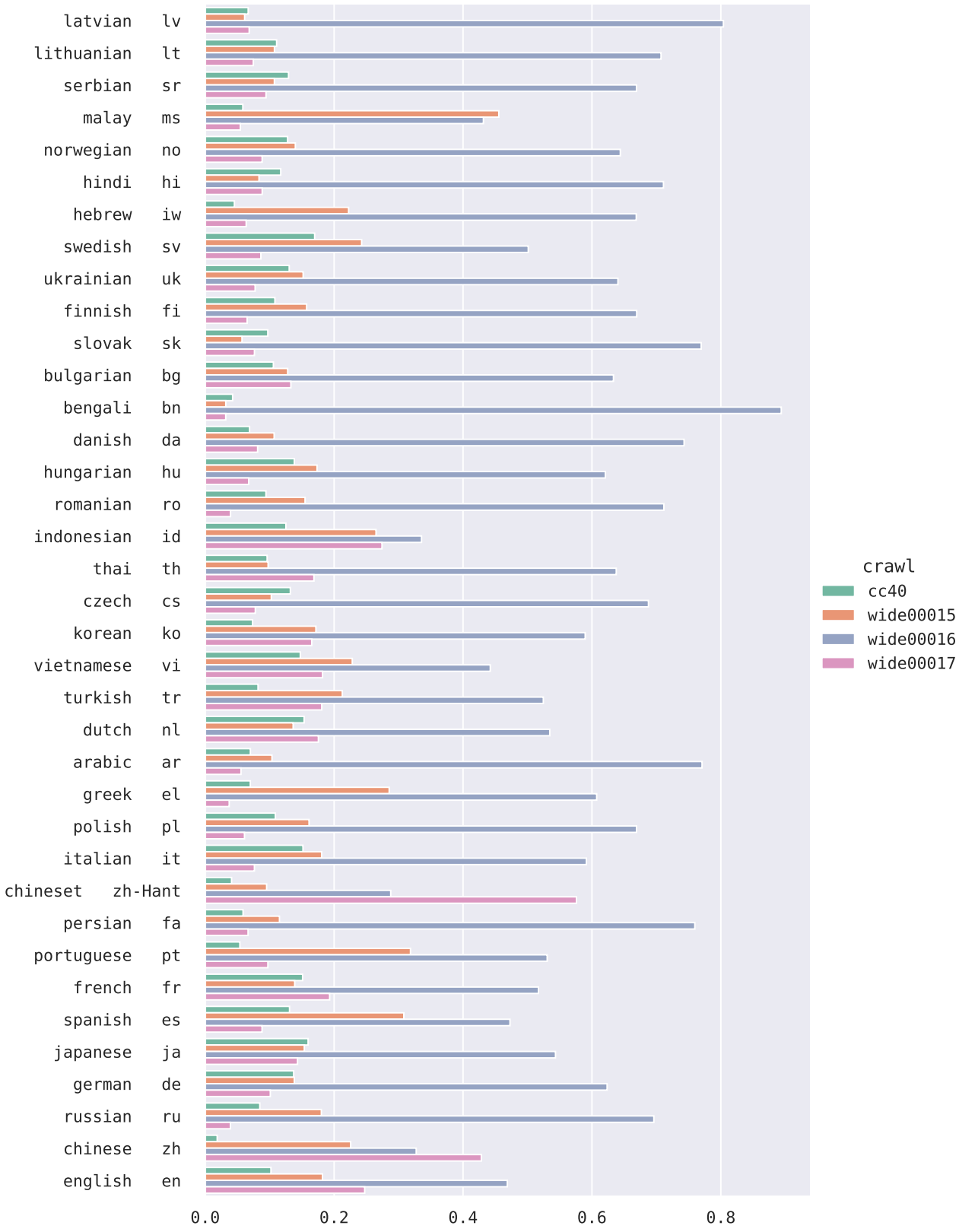


Figure 4.2: Proportions of text volume in bytes coming from each crawl, part 2.

4.3 Sharding

To deal with the amount of data and the imbalance between languages, the raw text records are recombined into equally sized batches per languages. Records are grouped in the same batches based on a hash derived from the domain name of the URL of the record, modulo 256, resulting in 256 shards, each containing one or many batches.

This grouping of records by domain is done to help with the identification of parallel texts (see section 5). The assumption is that pages that are translations of each other are likely to be using the same domain name. To account for translations being hosted on their own top-level domain, this part of the domain name is ignored when generating the hash. The top-level domain of the URL is determined using the Public Suffix List[4]. For monolingual text extraction, the division among shards is ignored.

4.4 Preliminary cleaning and formatting

This section describes the pre-processing steps applied to *monolingual* datasets only (steps specific for *bitexts* are described in Chapter 5).

At this stage, the data extracted from WARCs and then sharded has been processed with cleaning tools in order to perform fixes at character level and to enrich the data with additional metadata that will be useful to perform filtered versions for different applications. Character and encoding fixes have been applied using Monofixer[5]. Metadata includes language identification and fluency scores at paragraph level as produced by FastSpell[6] and Monocleaner[7] respectively.

Preliminary cleaning was run on the LUMI HPC cluster using SIGMA2 quota. To be able to process 124TB of compressed text and scale across many LUMI compute nodes, a whole new pipeline based on Slurm scripts was developed[8]. This pipeline performs the following processing steps:

1. Joins `url.gz` and `plain_text.gz` files coming from the previous step (sharding) into a tab-separated file, where each line contains a document URL, a text paragraph and a collection name. The tab-separated file is divided into batches of equal amount of uncompressed text to try to balance subsequent processing jobs.
2. Each batch file is processed on a single compute node and parallelised across the number of lines with GNU Parallel.
3. Every line, containing a paragraph, is processed by the character and encoding fixer Monofixer, including
   - Fixing mojibake (encoding errors).
   - Unescaping HTML entities.
   - Removing HTML tags.
4. Every paragraph also receives two new columns of metadata:

[4] https://publicsuffix.org
[5] https://github.com/bitextor/bifixer
[6] https://github.com/mbanon/fastspell
[7] https://github.com/bitextor/monocleaner
[8] https://github.com/hplt-project/monotextor-slurm

- Language as identified by FastSpell, which consists of FastText language identification which is then refined using Hunspell dictionaries for improved precision. This refinement consists in checking spelling errors with each Hunspell dictionary in a list of similar[9] languages to the one identified by FastText. The language whose dictionary produces less spelling errors, is the final prediction.
- Fluency score computed with a 7-gram modified Knesser-Ney character language model. Each language model (one per language) is trained on samples of about 200,000 sentences mostly coming from the monolingual part of OPUS corpora[1]. Only corpora coming from non-web-crawled data and languages not being automatically identified, are chosen. Data from Wikipedia dumps are used for languages not having enough data from OPUS. This fluency score can be used to estimate the 'quality' of paragraphs in the document, allowing to filter out noise which is detrimental for training language models. To obtain fluency score for each paragraph, the computed perplexity of the paragraph is normalized. To compute this normalization, three values have been calculated previously:
  - Upper limit: clean text average perplexity plus standard deviation.
  - Lower limit: noisy text average perplexity minus standard deviation.
  - Middle point: perplexity value in the middle between noisy and clean averages.

  Where clean text is the training set and noisy text is the same text with its characters scrambled. Then, during processing, each paragraph gets its perplexity value normalized according this values:
  - Values higher than the upper limit get a score of 1.
  - Values lower than the lower limit get a score of 0.
  - Values between the middle point and the upper limit, or between middle point and the lower limit, the perplexity value is normalized by the distance between the corresponding limit and the middle point.

5. Finally, each batch tab-separated file is converted to JSON-lines (JSONL) format, where each document is serialized as a JSON entity with the following fields:
   - Document language identified by CLD2 during the WARC extraction process.
   - Document URL.
   - Collection name.
   - All the paragraphs of the document joined together with end line separators.
   - For each paragraph, the language identified in step 4.
   - For each paragraph, the fluency score identified in step 4.

In addition to the processing explained above, some modifications to how languages are stored after WARC text extraction have been made:

- CLD2 uses old Hebrew '`iw`' language code, so it has been renamed to use the official '`he`'[10].

[9] https://github.com/mbanon/fastspell/blob/main/src/fastspell/config/similar.yaml
[10] https://www.loc.gov/standards/iso639-2/php/langcodes_name.php?iso_639_1=he

- Norwegian Bokmål is identified as '`no`' by CLD2, so it has been changed to '`nb`' to avoid possible confusions, as '`no`' may also refer to all the Norwegian variants, not only Bokmål.
- For consistency with the rest of the languages, where we are not separating by writing script, traditional and simplified Chinese, '`zh-Hant`' and '`zh-Hans`', have been joined into '`zh`'.
- Serbo-Croatian languages (Bosnian '`bs`', Croatian '`hr`' and Serbian '`sr`') have been merged under '`hbs`' code. Because of their mutual intelligibility, these languages are often mixed up with each other during language identification.

This leaves **75 languages** for this first public HPLT data release.

4.5 Statistics of the public monolingual data release

4.5.1 Per-language text sizes

Table 4.4 shows the total sizes of texts in each language for the publicly released data after text extraction and the preliminary cleaning steps above explained.

Language	Code	# Segments	# Words	# Characters	# Bytes	# Documents
Esperanto	eo	5.54e+07	2.96e+08	2.02e+09	2.07e+09	4.01e+05
Pashto	ps	6.03e+07	3.57e+08	1.76e+09	2.81e+09	3.13e+05
Tatar	tt	6.40e+07	2.61e+08	1.83e+09	2.97e+09	3.68e+05
Kyrgyz	ky	5.15e+07	2.63e+08	2.00e+09	3.48e+09	3.34e+05
Somali	so	8.02e+07	5.44e+08	3.61e+09	3.64e+09	6.76e+05
Welsh	cy	1.33e+08	6.24e+08	4.24e+09	4.28e+09	7.25e+05
Irish	ga	3.56e+08	1.43e+09	9.31e+09	9.97e+09	2.71e+06
Maltese	mt	3.61e+08	1.39e+09	9.61e+09	1.01e+10	9.26e+05
Basque	eu	3.38e+08	1.55e+09	1.13e+10	1.14e+10	2.29e+06
Norwegian Nynorsk	nn	3.99e+08	1.64e+09	1.14e+10	1.17e+10	1.85e+06
Uzbek	uz	2.18e+08	1.11e+09	8.71e+09	1.20e+10	1.37e+06
Gujarati	gu	2.28e+08	1.06e+09	6.79e+09	1.32e+10	9.15e+05
Swahili	sw	3.72e+08	2.10e+09	1.35e+10	1.36e+10	2.17e+06
Punjabi	pa	3.28e+08	1.34e+09	7.70e+09	1.60e+10	2.40e+06
Nepali	ne	2.63e+08	1.68e+09	1.09e+10	2.43e+10	2.11e+06
Icelandic	is	9.11e+08	3.48e+09	2.40e+10	2.61e+10	3.86e+06
Marathi	mr	3.56e+08	1.91e+09	1.27e+10	2.71e+10	1.64e+06
Mongolian	mn	4.89e+08	2.49e+09	1.73e+10	2.92e+10	2.50e+06
Kannada	kn	7.59e+08	2.10e+09	1.44e+10	2.94e+10	1.93e+06
Sinhalese	si	3.35e+08	2.35e+09	1.48e+10	3.00e+10	1.37e+06
Kazakh	kk	5.90e+08	2.53e+09	1.93e+10	3.38e+10	3.46e+06
Galician	gl	1.23e+09	5.02e+09	3.31e+10	3.40e+10	4.60e+06
Macedonian	mk	8.23e+08	3.24e+09	2.13e+10	3.47e+10	3.41e+06
Telugu	te	5.37e+08	2.43e+09	1.72e+10	3.59e+10	3.51e+06
Malayalam	ml	5.26e+08	2.67e+09	2.05e+10	4.32e+10	2.19e+06
Afrikaans	af	1.43e+09	6.09e+09	4.48e+10	4.53e+10	4.36e+06
Tagalog	tl	1.77e+09	7.07e+09	4.69e+10	4.74e+10	4.97e+06
Armenian	hy	1.16e+09	4.06e+09	2.95e+10	4.75e+10	3.99e+06
Urdu	ur	1.46e+09	6.59e+09	3.79e+10	5.05e+10	6.09e+06
Burmese	my	1.05e+09	4.41e+09	3.25e+10	6.13e+10	2.46e+06
Belarusian	be	9.26e+08	4.71e+09	4.00e+10	6.81e+10	3.91e+06
Albanian	sq	2.25e+09	1.15e+10	7.73e+10	8.09e+10	9.18e+06
Azerbaijani	az	2.86e+09	1.07e+10	8.07e+10	9.15e+10	1.04e+07
Latin	la	2.55e+09	1.44e+10	1.04e+11	1.04e+11	2.05e+07
Georgian	ka	2.16e+09	7.19e+09	5.22e+10	1.05e+11	6.46e+06

Tamil	ta	1.53e+09	9.30e+09	7.08e+10	1.57e+11	5.46e+06
Catalan	ca	5.28e+09	2.45e+10	1.56e+11	1.61e+11	2.42e+07
Slovenian	sl	5.60e+09	2.28e+10	1.57e+11	1.61e+11	1.92e+07
Estonian	et	6.30e+09	2.17e+10	1.66e+11	1.71e+11	2.12e+07
Latvian	lv	7.04e+09	2.74e+10	2.07e+11	2.21e+11	2.12e+07
Lithuanian	lt	8.48e+09	3.31e+10	2.41e+11	2.53e+11	3.24e+07
Malay	ms	1.16e+10	5.67e+10	3.45e+11	3.52e+11	2.83e+07
Norwegian Bokmål	nb	1.52e+10	5.55e+10	3.72e+11	3.79e+11	6.11e+07
Hindi	hi	8.88e+09	4.12e+10	2.46e+11	4.89e+11	3.37e+07
Serbo-Croatian	hbs	1.84e+10	6.94e+10	4.73e+11	5.19e+11	6.06e+07
Hebrew	he	1.50e+10	6.18e+10	3.69e+11	6.03e+11	4.62e+07
Ukrainian	uk	1.27e+10	5.30e+10	3.86e+11	6.56e+11	4.71e+07
Swedish	sv	2.36e+10	9.60e+10	6.39e+11	6.63e+11	9.61e+07
Finnish	fi	2.43e+10	8.06e+10	6.53e+11	6.74e+11	8.91e+07
Slovak	sk	2.89e+10	9.47e+10	6.49e+11	6.94e+11	9.04e+07
Bengali	bn	1.49e+10	4.62e+10	3.10e+11	7.26e+11	2.33e+07
Bulgarian	bg	1.58e+10	6.64e+10	4.48e+11	7.43e+11	5.83e+07
Danish	da	4.16e+10	1.40e+11	9.25e+11	9.44e+11	6.86e+08
Romanian	ro	3.29e+10	1.45e+11	1.10e+12	1.12e+12	1.04e+08
Hungarian	hu	3.76e+10	1.37e+11	1.03e+12	1.12e+12	1.37e+08
Czech	cs	5.23e+10	1.92e+11	1.26e+12	1.37e+12	1.84e+08
Thai	th	3.22e+10	7.86e+10	6.68e+11	1.39e+12	9.53e+07
Indonesian	id	2.95e+10	2.08e+11	1.44e+12	1.45e+12	1.26e+08
Korean	ko	5.56e+10	1.61e+11	8.34e+11	1.56e+12	2.48e+08
Vietnamese	vi	4.53e+10	2.87e+11	1.41e+12	1.73e+12	1.74e+08
Turkish	tr	5.03e+10	2.38e+11	1.68e+12	1.81e+12	2.15e+08
Dutch	nl	6.20e+10	2.50e+11	1.81e+12	1.83e+12	2.34e+08
Arabic	ar	4.01e+10	2.16e+11	1.37e+12	2.19e+12	1.97e+08
Greek	el	5.43e+10	2.48e+11	1.60e+12	2.57e+12	1.34e+08
Polish	pl	8.71e+10	3.66e+11	2.55e+12	2.65e+12	3.46e+08
Italian	it	8.82e+10	4.06e+11	2.76e+12	2.79e+12	3.37e+08
Persian	fa	6.31e+10	3.18e+11	1.74e+12	2.93e+12	1.89e+08
Portuguese	pt	1.26e+11	6.07e+11	3.79e+12	3.91e+12	4.48e+08
French	fr	1.73e+11	7.92e+11	5.12e+12	5.25e+12	6.60e+08
Spanish	es	1.62e+11	8.69e+11	5.38e+12	5.51e+12	6.72e+08
Japanese	ja	1.75e+11	3.05e+11	2.75e+12	5.97e+12	6.80e+08
German	de	2.22e+11	8.99e+11	6.72e+12	6.83e+12	1.02e+09
Russian	ru	4.34e+11	1.79e+12	1.33e+13	2.18e+13	1.53e+09
Chinese	zh	8.88e+11	1.79e+12	1.76e+13	4.00e+13	6.91e+09
English	en	2.20e+12	1.03e+13	6.78e+13	6.82e+13	1.27e+10
Total		5.40e+12	2.17e+13	1.51e+14	1.94e+14	2.89e+10

Table 4.3: Languages in the public data release: the number of segments (new line symbols), words (as defined by `wc(1)`), characters, bytes and documents. Ordered by size in bytes.

Language	**Code**	**# Segments**	**# Words**	**# Characters**	**# Bytes**	**# Documents**
Esperanto	eo	2.24e+07	1.53e+08	1.01e+09	1.03e+09	1.77e+05
Pashto	ps	2.64e+07	1.72e+08	8.82e+08	1.41e+09	1.43e+05
Welsh	cy	3.84e+07	2.34e+08	1.51e+09	1.53e+09	2.85e+05
Tatar	tt	2.64e+07	1.34e+08	9.65e+08	1.63e+09	1.72e+05
Somali	so	3.54e+07	2.54e+08	1.71e+09	1.73e+09	3.75e+05
Kyrgyz	ky	2.60e+07	1.53e+08	1.17e+09	2.06e+09	1.88e+05
Irish	ga	1.21e+08	5.20e+08	3.35e+09	3.58e+09	9.32e+05
Norwegian Nynorsk	nn	1.09e+08	6.16e+08	4.28e+09	4.38e+09	7.53e+05
Basque	eu	1.25e+08	6.60e+08	4.96e+09	5.02e+09	1.01e+06
Swahili	sw	1.22e+08	8.31e+08	5.47e+09	5.52e+09	9.84e+05
Gujarati	gu	6.10e+07	4.31e+08	2.66e+09	5.54e+09	4.55e+05

Maltese	mt	1.69e+08	8.19e+08	5.69e+09	6.01e+09	4.84e+05
Uzbek	uz	9.27e+07	5.56e+08	4.38e+09	6.12e+09	6.33e+05
Punjabi	pa	1.10e+08	5.23e+08	2.98e+09	6.26e+09	8.88e+05
Kannada	kn	1.03e+08	4.92e+08	3.57e+09	7.60e+09	5.58e+05
Galician	gl	2.22e+08	1.29e+09	8.34e+09	8.56e+09	1.79e+06
Sinhalese	si	9.54e+07	7.35e+08	4.78e+09	9.93e+09	5.64e+05
Icelandic	is	3.16e+08	1.56e+09	1.03e+10	1.13e+10	1.44e+06
Tagalog	tl	2.40e+08	1.63e+09	1.15e+10	1.16e+10	1.20e+06
Macedonian	mk	1.77e+08	1.07e+09	6.99e+09	1.18e+10	1.25e+06
Marathi	mr	1.24e+08	8.12e+08	5.41e+09	1.18e+10	8.57e+05
Mongolian	mn	1.48e+08	1.05e+09	7.28e+09	1.23e+10	1.06e+06
Afrikaans	af	2.02e+08	1.60e+09	1.32e+10	1.34e+10	1.37e+06
Kazakh	kk	2.18e+08	1.03e+09	7.85e+09	1.37e+10	1.43e+06
Nepali	ne	1.38e+08	9.67e+08	6.36e+09	1.49e+10	1.36e+06
Telugu	te	2.12e+08	1.03e+09	7.35e+09	1.55e+10	1.61e+06
Armenian	hy	2.58e+08	1.29e+09	9.52e+09	1.58e+10	1.36e+06
Urdu	ur	2.68e+08	2.02e+09	1.11e+10	1.61e+10	2.23e+06
Malayalam	ml	1.63e+08	9.17e+08	7.70e+09	1.72e+10	1.13e+06
Burmese	my	2.47e+08	1.14e+09	9.19e+09	1.96e+10	8.26e+05
Georgian	ka	3.58e+08	1.61e+09	1.20e+10	2.67e+10	1.67e+06
Albanian	sq	5.67e+08	3.66e+09	2.57e+10	2.72e+10	3.22e+06
Azerbaijani	az	4.99e+08	2.88e+09	2.39e+10	2.74e+10	3.00e+06
Belarusian	be	2.72e+08	1.93e+09	1.73e+10	3.04e+10	1.26e+06
Latin	la	4.06e+08	3.81e+09	3.30e+10	3.31e+10	4.81e+06
Latvian	lv	1.38e+09	5.85e+09	4.44e+10	4.74e+10	5.12e+06
Slovenian	sl	1.43e+09	6.72e+09	4.72e+10	4.82e+10	5.82e+06
Catalan	ca	1.16e+09	7.88e+09	4.94e+10	5.10e+10	7.79e+06
Tamil	ta	4.68e+08	2.94e+09	2.30e+10	5.13e+10	2.47e+06
Estonian	et	1.55e+09	6.57e+09	5.01e+10	5.16e+10	5.84e+06
Lithuanian	lt	1.66e+09	7.33e+09	5.42e+10	5.71e+10	7.40e+06
Bengali	bn	8.85e+08	4.86e+09	3.23e+10	7.31e+10	5.97e+06
Malay	ms	2.43e+09	1.34e+10	8.50e+10	8.66e+10	8.36e+06
Norwegian Bokmål	nb	2.98e+09	1.63e+10	1.06e+11	1.09e+11	1.46e+07
Slovak	sk	3.05e+09	1.42e+10	1.02e+11	1.10e+11	1.40e+07
Hebrew	he	2.68e+09	1.44e+10	8.65e+10	1.39e+11	1.12e+07
Serbo-Croatian	hbs	3.09e+09	1.79e+10	1.24e+11	1.42e+11	1.78e+07
Danish	da	4.58e+09	2.21e+10	1.53e+11	1.56e+11	2.36e+07
Finnish	fi	4.14e+09	1.99e+10	1.58e+11	1.64e+11	1.95e+07
Hindi	hi	2.38e+09	1.41e+10	8.39e+10	1.68e+11	1.14e+07
Bulgarian	bg	2.83e+09	1.53e+10	1.03e+11	1.71e+11	1.33e+07
Romanian	ro	4.57e+09	2.89e+10	1.88e+11	1.93e+11	2.49e+07
Swedish	sv	5.39e+09	2.98e+10	1.95e+11	2.02e+11	3.00e+07
Hungarian	hu	5.50e+09	2.80e+10	2.11e+11	2.28e+11	2.85e+07
Ukrainian	uk	3.18e+09	1.82e+10	1.34e+11	2.31e+11	1.79e+07
Czech	cs	6.96e+09	3.64e+10	2.48e+11	2.70e+11	3.86e+07
Korean	ko	8.72e+09	3.43e+10	1.71e+11	3.25e+11	4.45e+07
Vietnamese	vi	6.80e+09	5.92e+10	2.93e+11	3.62e+11	4.01e+07
Thai	th	4.84e+09	1.64e+10	1.68e+11	3.73e+11	2.95e+07
Dutch	nl	1.00e+10	5.59e+10	3.76e+11	3.79e+11	6.66e+07
Indonesian	id	7.43e+09	5.42e+10	3.78e+11	3.80e+11	4.58e+07
Turkish	tr	1.03e+10	6.49e+10	4.55e+11	4.93e+11	5.94e+07
Arabic	ar	8.31e+09	4.95e+10	3.02e+11	4.97e+11	4.66e+07
Greek	el	9.49e+09	4.99e+10	3.29e+11	5.31e+11	3.06e+07
Persian	fa	7.17e+09	5.75e+10	3.10e+11	5.31e+11	4.23e+07
Polish	pl	1.43e+10	7.63e+10	5.44e+11	5.67e+11	8.29e+07
Portuguese	pt	1.87e+10	1.22e+11	7.60e+11	7.86e+11	1.04e+08
Italian	it	2.28e+10	1.15e+11	8.22e+11	8.30e+11	9.65e+07
French	fr	2.63e+10	1.74e+11	1.14e+12	1.17e+12	1.76e+08
German	de	3.37e+10	1.91e+11	1.44e+12	1.47e+12	2.26e+08

Spanish	es	3.33e+10	2.39e+11	1.49e+12	1.53e+12	2.01e+08
Japanese	ja	4.09e+10	7.74e+10	8.60e+11	1.93e+12	2.19e+08
Russian	ru	6.58e+10	4.14e+11	3.09e+12	5.06e+12	3.97e+08
Chinese Simplified	zh	2.00e+11	4.83e+11	5.75e+12	1.34e+13	1.20e+09
English	en	3.87e+11	2.86e+12	2.03e+13	2.03e+13	1.78e+09
Total		9.84e+11	5.56e+12	4.15e+13	5.41e+13	5.25e+09

Table 4.4: Release 1.2, after deduplication: the number of segments (new line symbols), words (as defined by `wc(1)`), characters, bytes and documents. Ordered by size in bytes.

4.5.2 The distributions of fluency scores and segment lengths

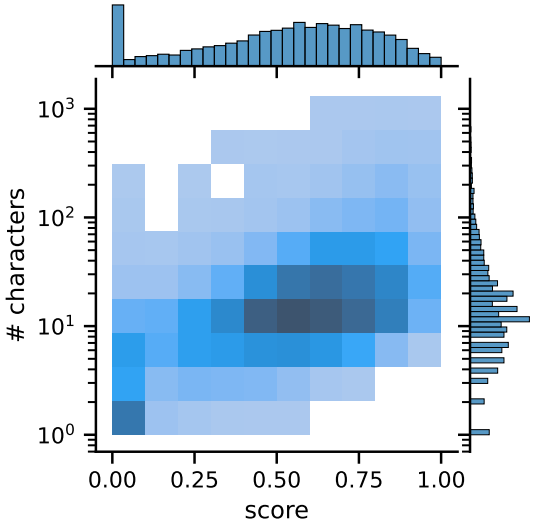


Figure 4.3: Joint distribution of fluency scores and lengths for a sample of segments from CC40 and WIDE17 stratified by language and crawl.

In order to study the distributions of fluency scores and segment lengths, a stratified random sample of documents was obtained from the released data coming from the WIDE17 and CC40 crawls. There are about 10K segments from each crawl for each language in the sample. Around 1.3% of segments are empty segments. Figure 4.3 shows the joint distribution of fluency scores and lengths for non-empty segments. Notably, most segments are quite short distributed around a few dozen characters. There is a significant portion of segments with fluency score of approximately zero, mostly shorter ones. The longer segments get fluency scores from the whole [0,1] interval centered around a score of 0.6. Fluency scores and segment lengths are correlated with the Spearman correlation coefficient of 0.54.

The distributions of fluency scores for each language and crawl separately are shown in Figure 4.4. All distributions are multi-modal with one mode located at zero. The amount of zero scores significantly vary from language to language, as well as the locations of other modes.

Figure 4.5 depicts the distributions of segment lengths depending on language and crawl. Except for a few languages, the differences are not so pronounced as in the case of fluency scores.

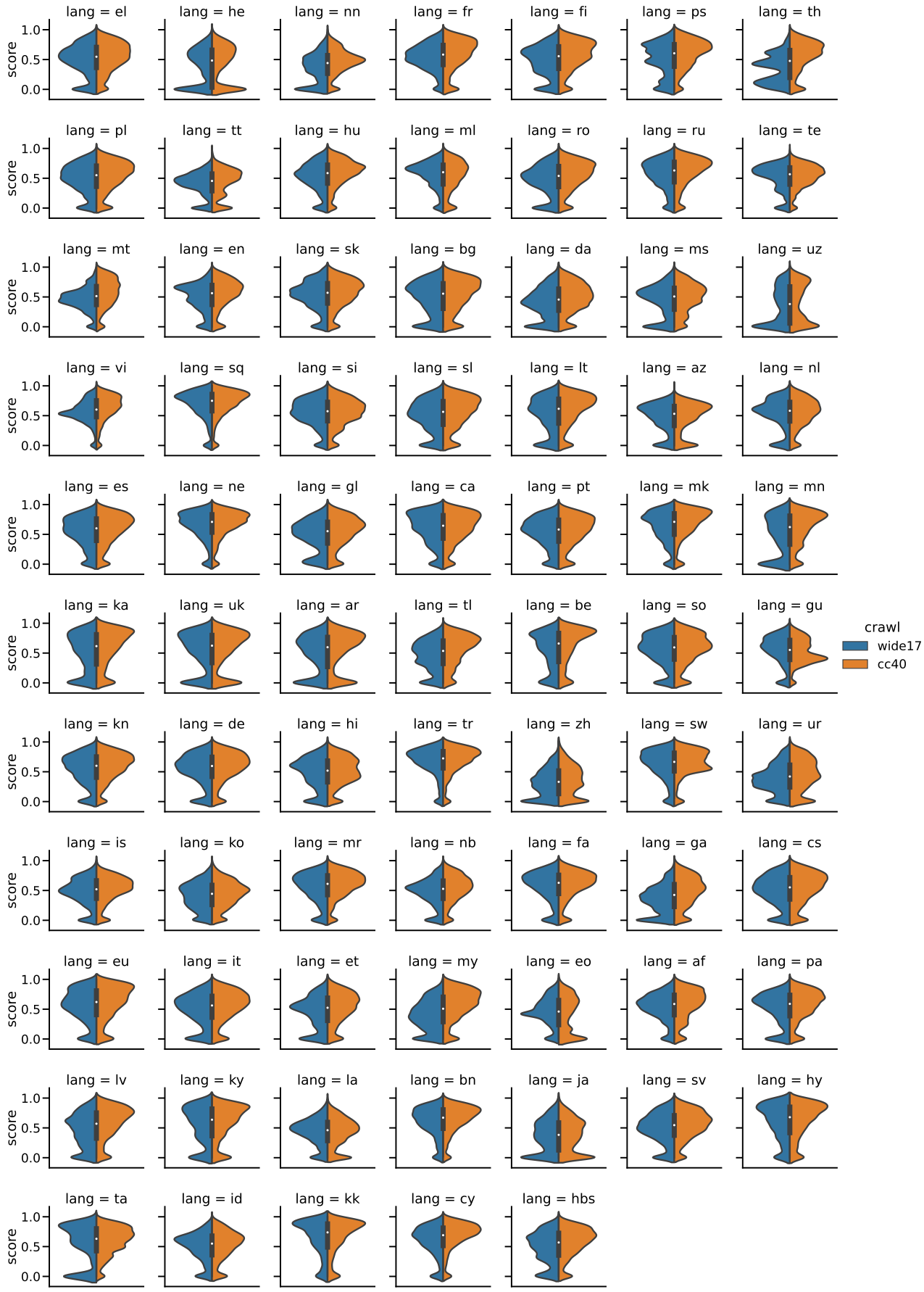


Figure 4.4: The distributions of fluency scores per language.

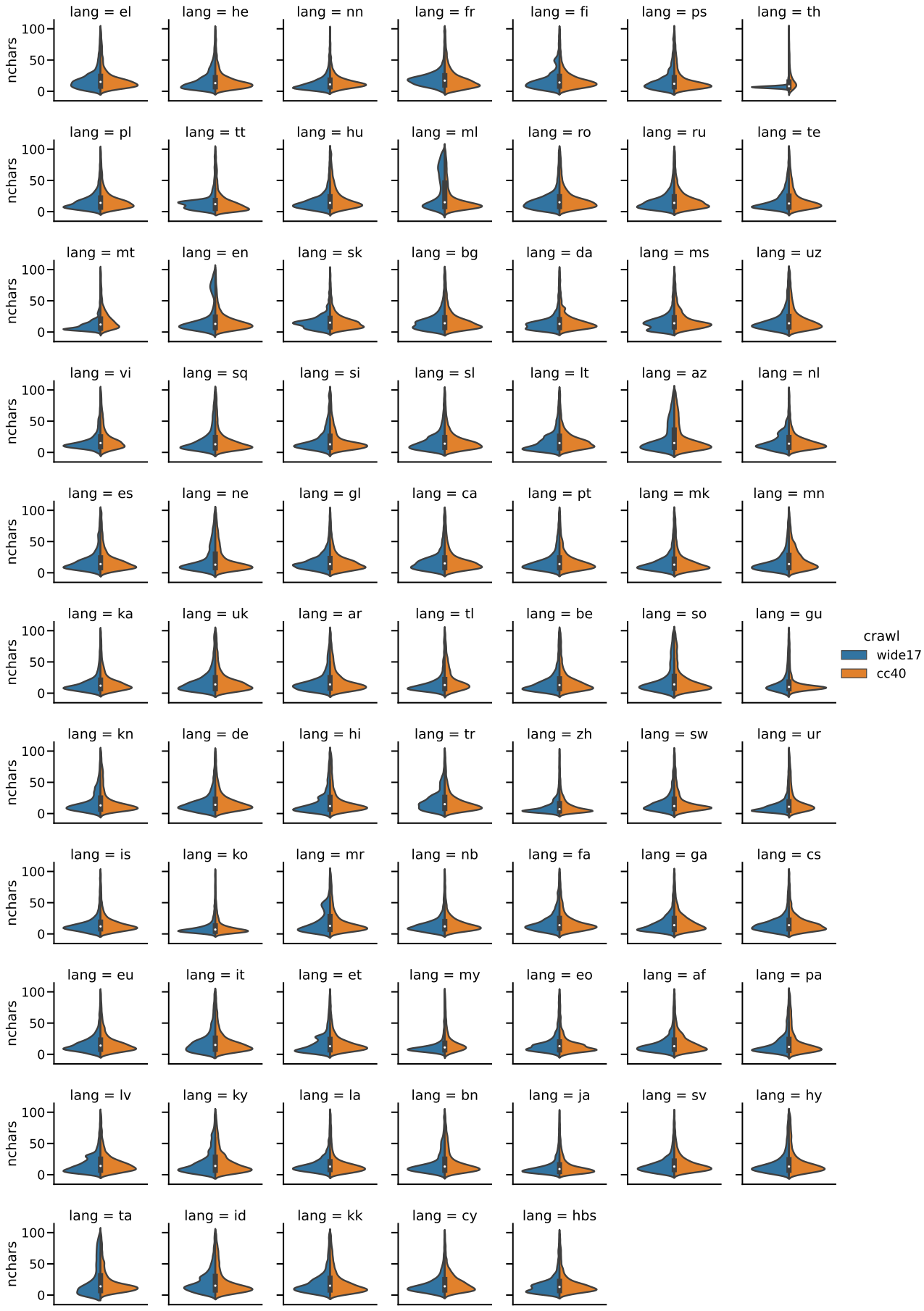


Figure 4.5: The distributions of segment lengths per language. For visual clarity, only segments shorter than 100 characters are taken, they cover 95% of all data.

5 Parallel Data

In this section, we describe newly created bitexts from the web crawls as well as imports and extensions of the OPUS collections done within the first year of the HPLT project. We start with the bitext extraction efforts before moving to the collection of additional parallel data sets and their integration in OPUS.

5.1 Bitext Extraction

The monolingual data sets extracted, split by language and sharded (see Sections 4 up to subsection 4.3) from the Internet Archive and CommonCrawl provide the input to the bitext extraction pipeline used for creating parallel corpora for this initial release. We rely on previous experience and tools from the ParaCrawl[1] and MaCoCu projects[2] and adjust tools and procedures from the Bitextor pipeline (see Section 5.2) according to the needs and languages in the HPLT setup.

In this release, we focus on English-centric data as we expect the largest potential outcome of parallel data from the alignment to English. Furthermore, the Bitextor pipeline relies on automatic document translation in one of the steps and the performance of translations into English is more reliable than translations into other languages, especially for lesser resourced languages. The inclusion of non-English-centric language pairs will be stressed in future releases. The initial relase covers 14 language pairs with a strong focus on lesser resourced languages also including a few non-European languages to increase the diversity of parallel data available for machine translation (MT) development.

Below, we first briefly present the extraction pipeline, the additional models we trained and released to apply it and then present some statistics about the parallel data in the current release. Thereafter, we describe extensions of OPUS with new imports and updates.

5.2 The Bitextor Pipeline

The bitext extraction pipeline is based on Bitextor.[3] Scripts[4] developed for ParaCrawl[2] were updated and used for scheduling and workflow automation.

Mining bilingual sentence pairs using this pipeline and English as one of the languages covers the following processing steps:

- Extract raw text from WARC archives, perform language identification and group records into shards and batches. (This step is shared with monolingual text extraction.)
- Split the documents into sentences using a language-specific sentence splitter.
- Translating sentences in languages different from English into English.
- Match English and translated documents using TF/IDF.

[1] https://www.paracrawl.eu/
[2] https://macocu.eu/
[3] https://github.com/bitextor/bitextor
[4] https://github.com/paracrawl/cirrus-scripts

- Match English and translated sentences in the matched documents using Bleualign[5], which produces the original untranslated sentence and the matched English sentence for any match.
- Fix encoding and orthographic issues with Bifixer[3], remove rule-based noisy sentence pairs using Bicleaner-hardrules and score sentence pairs using Bicleaner AI[4].

We adjusted and further developed the pipeline for the needs of HPLT on the LUMI supercomputer.[6] Downloading, storage of the WARC archives, and raw text extraction was done on separate clusters (NIRD, CESNET) to work around the limitations on the amount of data and number of files on LUMI. The size of the batches (in which raw text is stored) was increased to further reduce the number of individual files. MarianNMT was used for automatic translation, and adapted to work with the AMD GPUs available on LUMI.[7] Furthermore, we changed the document aligner code[8] to work with less memory to better handle the larger batch size. The LUMI supercomputer has more CPU cores available, but as a result less memory per CPU core.

Note that the current initial release does not cover the last cleaning step with dedicated Bicleaner models. It only covers preliminary fixes and cleaning of obvious noise provided by Bifixer and the Bicleaner-hardrules. Further cleaning and filtering will be applied in next releases to maximize the possibility to use alternative pre-processing pipelines when using the data. Users should, however, be aware of the risks of using the data sets as-is and we recommend to carefully look at the data before applying them in model training.

5.3 MT Models for Document Alignment

Document alignment in the Bitextor pipeline requires the translation of one language into the other in order to use efficient monolingual matching strategies to find parallel document in the vast space of extracted texts. This requires efficient translation models to enable computationally feasible jobs on the data we are looking at. Bitextor already supports a number of languages from prior work but the coverage is limited. OPUS-MT[9] provides additional resources in terms of pre-trained models that can be employed directly for translation or for distillation as explained below.

For this data release, we trained new efficient student MT models to enable the extraction of additional language pairs. We adopted larger transformer-based machine translation systems as teacher models and distilled knowledge from the teacher to train student models and improve efficiency. This technique helps the student model learn from the teacher model's insights, leading to a system of comparable quality, but improved throughput thanks to its smaller size. We trained two student models, i.e., base and tiny, for languages including `ar`, `eu`, `gl`, `hi`, `jp`, `sw`, `vi`, and `zh` (in both simplified and traditional scripts). We released the student models, as listed in Table 5.1, via CSC's Pouta cloud services.

More information about the procedures used to train models including scripts and configurations are available from GitHub.[10]

[5] https://github.com/bitextor/bleualign-cpp
[6] https://www.lumi-supercomputer.eu/
[7] https://github.com/hplt-project/lumi-marian
[8] https://github.com/hplt-project/document-aligner/pull/2
[9] https://github.com/Helsinki-NLP/OPUS-MT
[10] https://github.com/hplt-project/bitextor-mt-models

Language (Script)	Model Size	Link
ar	base	https://object.pouta.csc.fi/hplt_bitextor_models/ara_base.tar.gz
ar	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/ara_tiny.tar.gz
eu	base	https://object.pouta.csc.fi/hplt_bitextor_models/eus_base.zip
eu	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/eus_tiny.zip
gl	base	https://object.pouta.csc.fi/hplt_bitextor_models/gl-en_exported_base.zip
gl	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/gl-en_exported_tiny.zip
hi	base	https://object.pouta.csc.fi/hplt_bitextor_models/hin_base.tar.gz
hi	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/hin_tiny.tar.gz
jp	base	https://object.pouta.csc.fi/hplt_bitextor_models/jpn-eng.base.zip
jp	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/jpn-eng.tiny.zip
sw	base	https://object.pouta.csc.fi/hplt_bitextor_models/sw-en_exported_base.zip
sw	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/sw-en_exported_tiny.zip
vi	base	https://object.pouta.csc.fi/hplt_bitextor_models/vie-eng.base.zip
vi	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/vie-eng.tiny.zip
zh_Hans	base	https://object.pouta.csc.fi/hplt_bitextor_models/zho_hans.base.zip
zh_Hans	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/zho_hans.tiny.zip
zh_Hant	base	https://object.pouta.csc.fi/hplt_bitextor_models/zho_hant.base.zip
zh_Hant	tiny	https://object.pouta.csc.fi/hplt_bitextor_models/zho_hant.tiny.zip

Table 5.1: Released student machine translation models

5.4 Bicleaner Models for Data Filtering

Although we did not apply the final cleaning step to this data release, all needed Bicleaner AI models were trained and are available to download[11] for the language pairs that we include in D2.1. We have increased the total amount of language pairs available from 36 to 45[12], also including many changes and improvements to the tool since version 1.0.1 made for Paracrawl[13].

5.5 Extracted Bitexts

The newly created bitexts come in the form of aligned documents with sentences linked to each other. After sentence alignment, we apply several straightforward fixes to the text and hard rules filtering including automatic language identification reassessment to extract reasonable sentence pairs from the raw alignment. For this initial release, no additional filtering, cleaning nor de-duplication is applied. Hence, all data sets still include a substantial proportion of noise and repetition.

Below, we provide further statistics of the material we have extracted so far in table 5.2. We show the size of each bitext after applying fixes and rule-based filtering including information about the proportion of the original aligned data that is retained (percentage in the bytes column). The substantial reduction of this filtering step shows the impact of that procedure on this kind of noisy data coming from web crawls. From manual inspection, we can verify that this step is essential to remove large portions of the noise that naturally appears in web crawled data sets such as untranslated text and other kind of noise that leads to misalignments and off-target segments.

[11] https://github.com/bitextor/bicleaner-ai#download-a-model
[12] https://huggingface.co/models?other=bicleaner-ai
[13] https://github.com/bitextor/bicleaner-ai/blob/v2.3.2/CHANGELOG.md

Language	Code	# Segments	# Words	# Characters	# Bytes	# Documents
Norwegian	nn	9.34e+05	9.90e+06	6.08e+07	6.10e+07 (6.7%)	1.18e+05
Bosnian	bs	1.95e+06	1.71e+07	1.07e+08	1.07e+08 (10.8%)	2.67e+05
Basque	eu	4.55e+06	4.77e+07	2.96e+08	2.97e+08 (37.7%)	2.92e+05
Galician	gl	7.45e+06	6.50e+07	4.02e+08	4.03e+08 (21.0%)	5.43e+05
Serbian	sr	9.05e+06	7.08e+07	4.56e+08	4.57e+08 (2.0%)	2.12e+06
Gaelic	ga	1.95e+07	1.85e+08	1.12e+09	1.12e+09 (31.1%)	1.82e+06
Maltese	mt	1.35e+07	1.82e+08	1.15e+09	1.16e+09 (19.9%)	1.15e+06
Albanian	sq	2.18e+07	1.97e+08	1.18e+09	1.18e+09 (8.6%)	2.42e+06
Macedonian	mk	2.68e+07	2.86e+08	1.73e+09	1.74e+09 (50.2%)	1.06e+06
Icelandic	is	3.68e+07	3.44e+08	2.06e+09	2.06e+09 (34.6%)	1.37e+06
Arabic	ar	6.08e+07	5.65e+08	3.53e+09	3.54e+09 (54.0%)	2.42e+06
Estonian	et	1.32e+08	1.17e+09	7.12e+09	7.14e+09 (25.8%)	6.41e+06
Hindi	hi	1.70e+08	1.44e+09	8.90e+09	8.92e+09 (26.3%)	9.43e+06
Croatian	hr	1.62e+08	1.53e+09	9.48e+09	9.50e+09 (31.5%)	6.93e+06
Total		6.67e+08	6.12e+09	3.76e+10	3.77e+10 (23.8%)	3.63e+07

Table 5.2: Statistics on the extracted bitexts: the number of segments, words, characters, bytes (in parentheses, percentage of remaining bytes after applying hard-rules) and documents. All statistics are measured from the English side of each language pair. Ordered by size in bytes.

The focus of our efforts was set on setting a stable pipeline and extracting data for under-resourced languages. The list is still very limited and we will extend the coverage throughout the project. This initial release contains over 660 million segments and more than 6 billion words in 14 language pairs.

Figure 5.1 shows the proportions of data coming from each crawl, for each language pair separately. For some languages, not all collections were processed. Similar to the monolingual data, most aligned text comes from the WIDE16 collection, except in the case of Basque (eu), where it comes primarily from CC40.

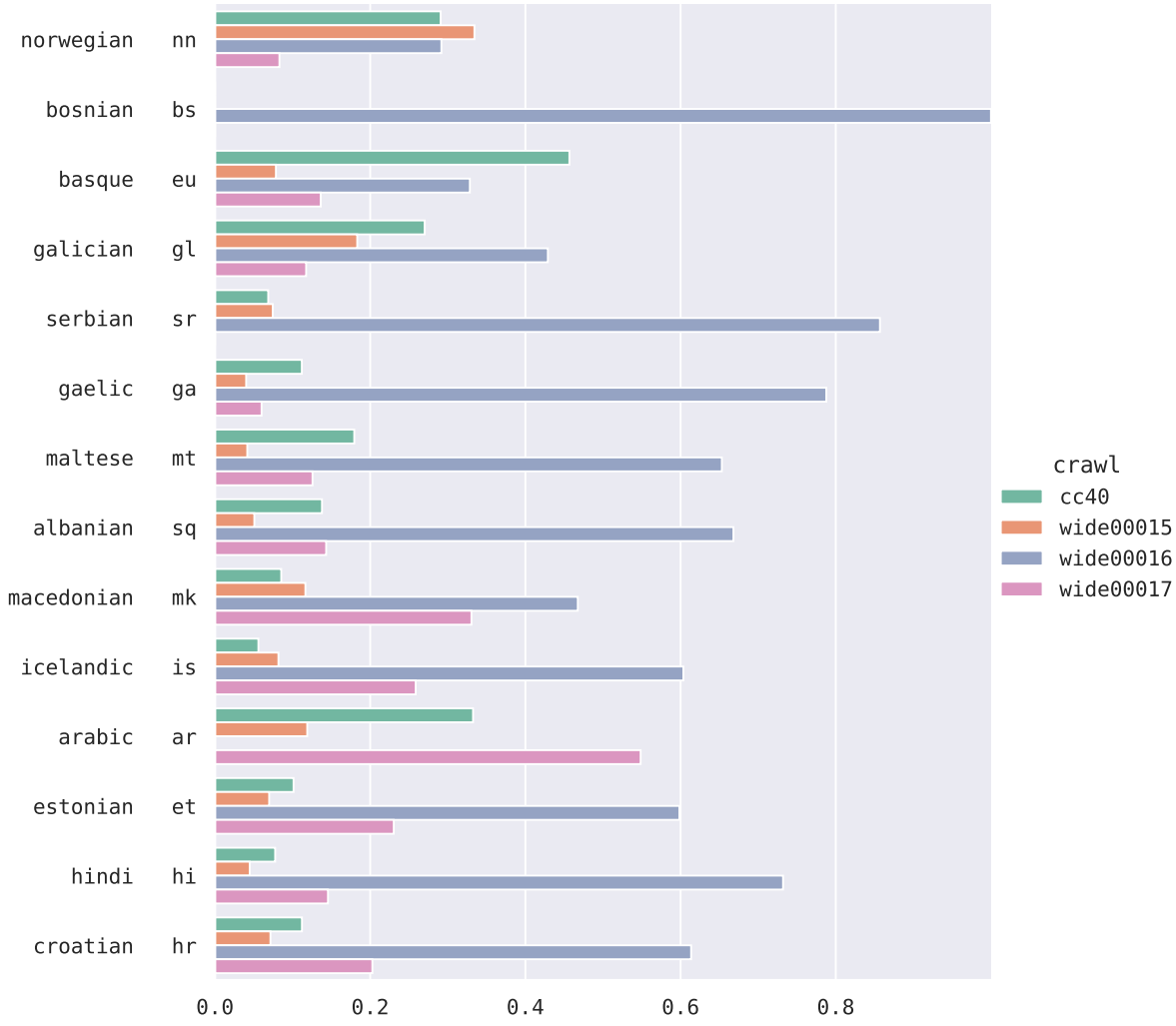


Figure 5.1: Proportions of aligned text sizes coming from each crawl.

5.6 Further Bitext Collection

During the first year of HPLT, we also imported a large number of additional resources to the parallel data collection in OPUS. Together with the import, we also made a substantial effort to streamline import procedures and to improve data structures and consistencies. Essential information and metadata about releases is now available in a version-controlled repository at GitHub[14] and import procedures are stored in another git repository.[15]. Besides version control and transparency, GitHub also offers a better way of handling issues using trackers and development functionalities available at the platform. In addition, we can more easily collect information about new resources that can be queued for import into the collection.

Part of the effort was a systematic import of resources published at the ELRC-Share repository[16]. The

[14] https://github.com/Helsinki-NLP/OPUS
[15] https://github.com/Helsinki-NLP/OPUS-ingest/
[16] https://lr-coordination.eu/

language	files	tokens	sentences	bg	cs	da	de	el	en	es	et	fi	fr	ga	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	1	0.4M	13.2k		12.8k	12.8k	12.7k	12.8k	13.2k	12.7k	12.7k	12.6k	12.6k	0.4k	12.5k	12.6k	12.7k	12.2k	12.6k	2.9k	12.8k	12.5k	12.7k	12.7k	12.6k	12.8k	12.7k
cs	1	0.3M	13.1k	12.8k		12.8k	12.8k	12.8k	13.2k	12.7k	12.7k	12.7k	12.6k	0.4k	12.5k	12.6k	12.7k	12.3k	12.6k	2.9k	12.9k	12.5k	12.7k	12.7k	12.7k	12.9k	12.7k
da	1	0.4M	13.2k	12.8k	12.9k		12.9k	12.8k	13.2k	12.8k	12.8k	12.7k	12.7k	0.4k	12.6k	12.7k	12.8k	12.3k	12.6k	2.9k	12.9k	12.6k	12.8k	12.8k	12.7k	12.9k	12.8k
de	1	0.4M	13.4k	12.8k	12.8k	12.9k		12.7k	13.3k	12.7k	12.7k	12.6k	12.8k	0.4k	12.5k	12.6k	12.7k	12.2k	12.5k	2.9k	12.8k	12.5k	12.7k	12.7k	12.6k	12.8k	12.7k
el	1	0.4M	13.1k	12.8k	12.8k	12.9k	12.8k		13.1k	12.7k	12.6k	12.6k	12.6k	0.4k	12.5k	12.6k	12.7k	12.2k	12.6k	2.9k	12.8k	12.4k	12.7k	12.7k	12.6k	12.8k	12.6k
en	1	0.4M	13.9k	13.2k	13.2k	13.2k	13.3k	13.1k		13.2k	13.0k	12.9k	13.2k	0.4k	12.8k	13.0k	13.1k	12.6k	13.0k	3.1k	13.2k	12.8k	13.1k	13.2k	12.9k	13.2k	13.1k
es	1	0.4M	13.2k	12.8k	12.8k	12.9k	12.7k	12.7k	13.2k		12.7k	12.6k	12.8k	0.4k	12.5k	12.6k	12.8k	12.2k	12.5k	2.9k	12.9k	12.5k	12.9k	12.8k	12.6k	12.8k	12.7k
et	1	0.3M	12.9k	12.7k	12.8k	12.8k	12.7k	12.7k	13.0k	12.7k		12.6k	12.5k	0.4k	12.4k	12.5k	12.6k	12.2k	12.5k	2.9k	12.7k	12.4k	12.6k	12.6k	12.5k	12.7k	12.6k
fi	1	0.3M	12.9k	12.6k	12.7k	12.7k	12.7k	12.6k	12.9k	12.6k	12.6k		12.4k	0.4k	12.3k	12.5k	12.5k	12.1k	12.4k	2.9k	12.6k	12.3k	12.5k	12.5k	12.4k	12.7k	12.6k
fr	1	0.4M	13.1k	12.7k	12.7k	12.8k	12.8k	12.6k	13.2k	12.8k	12.6k	12.5k		0.4k	12.3k	12.4k	12.6k	12.0k	12.4k	2.8k	12.7k	12.3k	12.7k	12.6k	12.4k	12.6k	12.5k
ga	1	7.4k	0.5k	0.4k	0.4k	0.4k	0.4k	0.4k	0.4k	0.4k	0.4k	0.4k	0.4k		0.4k	0.4k	0.4k	0.4k	0.4k	0.3k	0.4k	0.4k	0.4k	0.4k	0.3k	0.4k	0.4k
hr	1	0.3M	13.7k	12.5k	12.5k	12.6k	12.5k	12.5k	12.8k	12.5k	12.4k	12.3k	12.3k	0.4k		12.4k	12.4k	12.0k	12.4k	2.9k	12.5k	12.2k	12.4k	12.5k	12.4k	12.6k	12.4k
hu	1	0.3M	13.0k	12.7k	12.7k	12.7k	12.6k	12.6k	13.0k	12.6k	12.6k	12.5k	12.4k	0.4k	12.5k		12.6k	12.2k	12.6k	2.9k	12.7k	12.4k	12.6k	12.7k	12.5k	12.8k	12.6k
it	1	0.4M	13.1k	12.7k	12.8k	12.8k	12.7k	12.7k	13.1k	12.8k	12.6k	12.6k	12.6k	0.4k	12.5k	12.6k		12.1k	12.4k	2.9k	12.8k	12.4k	12.7k	12.6k	12.5k	12.7k	12.6k
lt	1	0.3M	12.7k	12.3k	12.3k	12.3k	12.3k	12.2k	12.6k	12.3k	12.2k	12.1k	12.1k	0.4k	12.0k	12.3k	12.2k		12.1k	2.8k	12.3k	12.0k	12.2k	12.2k	12.1k	12.3k	12.2k
lv	1	0.3M	13.0k	12.6k	12.6k	12.6k	12.6k	12.6k	13.0k	12.6k	12.5k	12.4k	12.4k	0.4k	12.5k	12.6k	12.5k	12.2k		2.9k	12.8k	12.4k	12.6k	12.7k	12.6k	12.8k	12.6k
mt	1	85.0k	3.1k	2.9k	2.9k	2.9k	3.0k	2.9k	3.1k	2.9k	2.9k	2.9k	2.8k	0.3k	2.9k	2.9k	2.9k	2.8k	2.9k		2.9k	2.9k	3.0k	3.0k	2.9k	3.0k	2.9k
nl	1	0.4M	13.2k	12.8k	12.9k	13.0k	12.9k	12.8k	13.2k	12.9k	12.8k	12.7k	12.7k	0.4k	12.6k	12.8k	12.8k	12.3k	12.8k	2.9k		12.5k	12.8k	12.7k	12.6k	12.9k	12.7k
pl	1	0.3M	13.0k	12.5k	12.6k	12.6k	12.5k	12.5k	12.8k	12.5k	12.4k	12.3k	12.3k	0.4k	12.2k	12.4k	12.4k	12.0k	12.5k	2.9k	12.5k		12.4k	12.3k	12.3k	12.5k	12.3k
pt	1	0.4M	13.1k	12.8k	12.8k	12.9k	12.8k	12.7k	13.1k	12.9k	12.6k	12.6k	12.7k	0.4k	12.5k	12.7k	12.7k	12.2k	12.7k	3.0k	12.8k	12.4k		12.8k	12.6k	12.8k	12.7k
ro	1	0.4M	13.7k	12.7k	12.8k	12.8k	12.7k	12.7k	13.2k	12.8k	12.6k	12.6k	12.6k	0.4k	12.5k	12.7k	12.6k	12.2k	12.7k	3.0k	12.7k	12.4k	12.8k		12.6k	12.9k	12.7k
sk	1	0.3M	12.9k	12.7k	12.7k	12.7k	12.6k	12.6k	12.9k	12.6k	12.5k	12.5k	12.4k	0.3k	12.4k	12.6k	12.5k	12.2k	12.6k	2.9k	12.6k	12.3k	12.6k	12.7k		12.6k	12.5k
sl	1	0.3M	13.2k	12.9k	12.9k	13.0k	12.9k	12.8k	13.2k	12.9k	12.8k	12.7k	12.7k	0.4k	12.6k	12.8k	12.8k	12.3k	12.8k	3.0k	12.9k	12.5k	12.8k	12.9k	12.6k		12.8k
sv	1	0.3M	13.1k	12.7k	12.8k	12.9k	12.7k	12.7k	13.1k	12.7k	12.7k	12.6k	12.5k	0.4k	12.5k	12.6k	12.6k	12.2k	12.7k	2.9k	12.8k	12.4k	12.7k	12.7k	12.5k	12.8k	

Figure 5.2: An example of a merged multilingual resource made out of individual bilingual ELRC packages coming from EU publications in the medical domain. The table shows the size of downloadable bitexts in plain text format (upper triangle) and TMX format (lower triangle).

resources in that collection are valuable but not straightforward to be used by the NLP community. There are various inconsistencies, overlaps and download complications that prevent the use of the bitexts in a streamlined way. OPUS now contains the majority of the public bitexts released through ELRC-Share in a unified format that can be accessed with the same convenient tools as all other resources in OPUS. Altogether, we have now 960 resources compiled from data released at ELRC-Share. Metadata can be used to trace back the original release and information available at ELRC-Share.

One of the peculiarities of ELRC releases is the division of multilingual resources into smaller bilingual sub-corpora. This is not only annoying but also prevents to see the full multilingual picture of a resource. Therefore, in addition to the plain import of ELRC-Share packages, we also create packages that merge such resources into one multilingual release. Examples are the COVID-19 ANTIBIOTIC datasets[17] and bilingual corpora from the Publications Office of the EU on the medical domain.[18] Note, that many ELRC-Share data sets are really small as well and the number of released packages does not necessarily reflect the amount of actual data that can be retrieved from the source (see Figure 5.2 for example for parallel data from EU publications – each bitext contains less than 15,000 translation units).

Another example for a multilingual corpus that has been split up into several bilingual releases is based on publications by the European Medicines Agency (EMA). The collection comes for historical reasons with the name EMEA and bilingual ELRC releases have been merged in OPUS as well.[19]. An earlier version of that data set was originally published in OPUS[20] and, hence, there is a huge overlap between the old and the new release. This is by far no exception and EMEA is the most straightforward case, where such overlap is very apparent. But, many data sets at ELRC-Share are compilations from existing resources and release names and information do not necessarily reveal the

[17] Multilingual OPUS version: https://opus.nlpl.eu/ELRC-antibiotic.php
[18] Multilingual OPUS version: https://opus.nlpl.eu/ELRC-EU_publications.php
[19] https://opus.nlpl.eu/ELRC-EMEA.php
[20] https://opus.nlpl.eu/EMEA.php

https://github.com/Helsinki-NLP/OPUS/blob/main/corpus/ELRC-EMEA/v1/overlaps/en-it.tsv

Preview Code Blame 100 lines (100 loc) · 5.31 KB

Search this file

	corpus A	release A	corpus B	release B	size A	size B	A∩B	A∩B/A	A∩B/B
2	ELRC-EMEA	v1	ELRC-2710-EMEA	v1	771346	771346	771323	99.99	99.99
3	ELRC-EMEA	v1	TildeMODEL	v2018	771346	3693634	186039	24.11	5.03
4	ELRC-EMEA	v1	EMEA	v3	771346	352671	51242	6.64	14.52
5	ELRC-EMEA	v1	CCMatrix	v1	771346	145080018	31664	4.10	.02
6	ELRC-EMEA	v1	ParaCrawl	v9	771346	96975992	13530	1.75	.01
7	ELRC-EMEA	v1	CCAligned	v1	771346	14341296	7584	.98	.05
8	ELRC-EMEA	v1	LinguaTools-WikiTitles	v2014	771346	11048442	1521	.19	.01
9	ELRC-EMEA	v1	DGT	v2019	771346	3638405	1328	.17	.03

Figure 5.3: A table listing overlaps between bitexts in different parallel corpora in OPUS. The example shows German-Italian taking ELRC-EMEA as the source for comparison. The last two columns are given in percentages.

relations to other public public data sets. This creates a huge problem for MT modelling as training data becomes increasingly infected by unnatural duplication and repetitions.

https://github.com/Helsinki-NLP/OPUS/blob/main/corpus/ELRC-EMEA/v1/overlaps/de.tsv

Preview Code Blame 178 lines (178 loc) · 9.92 KB

Search this file

	corpus A	release A	corpus B	release B	size A	size B	A∩B	A∩B/A	A∩B/B
2	ELRC-EMEA	v1	ELRC-2714-EMEA	v1	713209	713209	713177	99.99	99.99
3	ELRC-EMEA	v1	ELRC_2682	v1	713209	589704	589677	82.67	99.99
4	ELRC-EMEA	v1	TildeMODEL	v2018	713209	5030606	263320	36.92	5.23
5	ELRC-EMEA	v1	CCMatrix	v1	713209	534914327	101616	14.24	.01
6	ELRC-EMEA	v1	EMEA	v3	713209	349705	80874	11.33	23.12

Figure 5.4: Overlaps in terms of monolingual data counting identical sentences in different resources (here sentences in German in various data sets derived from EMA).

In order to make information about overlaps available to guide data selection decisions, we created procedures to systematically measure the amount of data that appears in several resources. For this, we compute the exact matches of aligned sentences in all pairs of bitexts and store tables that provide the counts and overlap percentages. Figure 5.3 shows and example for German-Italian and data extracted from EMA. Using this information, we can easily see that the bitext in the merged multilingual data set is more or less identical to the bilingual release (some minor differences seem to be caused by the merging process, which is also an interesting insight) and that there are major overlaps to at least two other resources indexed by OPUS (TildeModel and EMEA). CCMatrix also contains a substantial amount of identical sentence pairs. Note, that this count only includes exact matches of the entire

translation unit (excluding spaces). Small differences that may appear due to minor changes in pre-processing pipelines are not captured by those scores. Systematically measuring the overlaps is quite expensive with growing data sets but we consider the detection of near-duplicates in the future as well. Similarly, we can also measure the overlap in terms of monolingual data counting identical sentences. Figure 5.4 illustrates an example for German. This is useful to give a more complete picture on the situation as translation units are also effected by minor differences in alignment.

Corpus	Version	sentence count	TU's
ALT	v2019	36,292	18,086
Anuvaad	v1	37,015,993	16,037,705
DOGC	v3	17,577,583	8,472,148
ECDC	v2016-03-16	65,844	665,432
FFR	v2	164,268	77,978
GoURMET	v2	5,227,445	1,818,561
IITB	v2.0	3,306,533	1,331,242
JESC	v2019-12-05	5,598,453	2,797,388
JParaCrawl	v3.0	53,468,899	21,974,690
Joshua-IPC	v1	410,863	1,091,838
KFTT	v1.0	888,331	428,915
LinguaTools-WikiTitles	v2014	978,553,420	474,682,581
NeuLab-TedTalks	v1	6,448,186	72,354,431
Nunavut Hansard	v3.0.1	2,739,661	771,115
ParIce	v1	4,340,994	1,496,644
Samanantar	v0.2	101,397,540	49,745,409
StanfordNLP-NMT	v1.0	42,375,047	11,207,362
Tatoeba	v2023-04-12	11,561,057	8,513,233
XLEnt	v1.2	185,359,257	866,940,560
pmindia	v1	859,313	1,696,081
wikimedia	v20230407	53,034,210	21,385,243

Table 5.3: This table shows resources that have been imported to OPUS within HPLT excluding bitexts released at the ELRC-Share repository. Sentence counts refer to the total count of sentences or sentences fragments across all languages in each corpus and TU's refer to unique translation units per corpus. Versioning depends on the corpus and some of the resources above refer to updates of already existing corpora (XLEnt, GoURMET, Tatoeba and wikimedia).

Besides ELRC, there are various other resources that became available in recent years. Table 5.3 lists corpora that we have integrated and updated since the beginning of HPLT. Those resources are in general much bigger than ELRC packages and also contain genuinely new data points and additional language pairs, valuable resources to be included in our collection. Similarly to ELRC imports, we convert those data sets to the unified OPUS format to make them readily available for NLP research and MT training.

Altogether, we added 981 resources to OPUS covering 4.2 billion translation units. More information about individual releases and data sets is available from the GitHub repository and our website.

6 Packaging and Release Information

For releasing the data, we follow the principles of ParaCrawl[1] using the following licensing scheme:

- We do not own any of the text from which these data has been extracted.
- We license the actual packaging of these parallel data under the Creative Commons CC0 license ("no rights reserved").

This scheme will be complemented with a straightforward notice and take down policy stating that we remove affected sources with legitimate requests including a notice about the material that is claimed to be infringing and information reasonably sufficient to allow us to locate the material, as follows:

Notice and take down policy

Notice: Should you consider that our data contains material that is owned by you and should therefore not be reproduced here, please:

- Clearly identify yourself, with detailed contact data such as an address, telephone number or email address at which you can be contacted.
- Clearly identify the copyrighted work claimed to be infringed.
- Clearly identify the material that is claimed to be infringing and information reasonably sufficient to allow us to locate the material.
- And contact the HPLT project using the following email: hplt-datasets at ufal.mff.cuni.cz).

Take down: We will comply to legitimate requests by removing the affected sources from the next release of the corpus.

Data packaging

We ship the data sets as compressed packages per language in the monolingual case and per language pair for parallel data:

- Monolingual data comes in JSON format compressed with `zstd` and a plain text file linking to URLs of the archive. Downloads are possible with simple HTTPS requests using `wget` or the like.
- Bitexts come in TAB-separated file format (tsv) where, besides source and target segments, other metadata info such as source URLs are provided. The bitexts will also appear in OPUS as a new release of ParaCrawl after applying additional filtering using bicleaner-ai scores. After the integration into OPUS, all parallel data sets will become available through the unified OPUS API.

All downloads are available from https://hplt-project.org/datasets/. Metadata with direct links to data are also available as permanent metadata record in the LINDAT/CLARIAH-CZ repository at https://lindat.cz/repository.

[1] https://www.paracrawl.eu/

Bibliography

[1] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[2] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: https://aclanthology.org/2020.acl-main.417

[3] G. Ramírez-Sánchez, J. Zaragoza-Bernabeu, M. Bañón, and S. Ortiz-Rojas, "Bifixer and bicleaner: two open-source tools to clean your parallel data." in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, November 2020, pp. 291–298.

[4] J. "Zaragoza-Bernabeu, G. Ramírez-Sánchez, M. Bañón, and S. Ortiz Rojas, ""bicleaner AI: Bicleaner goes neural"," in *"Proceedings of the Thirteenth Language Resources and Evaluation Conference"*. "Marseille, France": "European Language Resources Association", Jun. "2022", pp. "824–831". [Online]. Available: "https://aclanthology.org/2022.lrec-1.87"