HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
dyu_Latn.jsonl.tsv	11/27/2024	Dyula (dyu)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,390	24,558	20,698 (84.28 %)	1.5M	5.7 MB	5,529,102

Top 10 domains

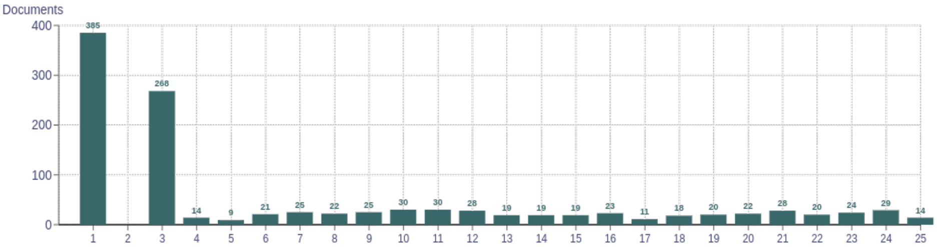
Domain	Docs	% of total
bible.is	646	46.47
bibles.org	431	31.01
jw.org	299	21.51
omniglot.com	4	0.29
bible.com	3	0.22
watchtower.org	3	0.22
gospelgo.com	2	0.14
twr360.org	1	0.07
reunion.com	1	0.07

Top 10 TLDs

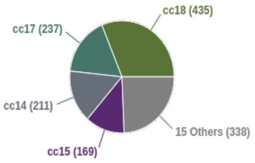
Domain	Docs	% of total
org	734	52.81
is	646	46.47
com	10	0.72

Documents size (in segments)

<= 25 segments **80.79%** (1.1K documents)
> 25 segments **19.21%** (267 documents)

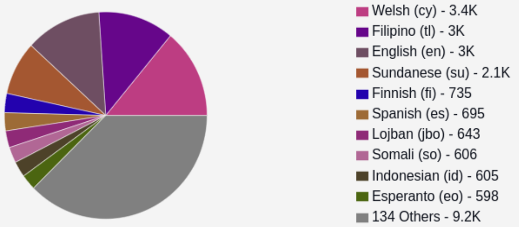


Documents by collection



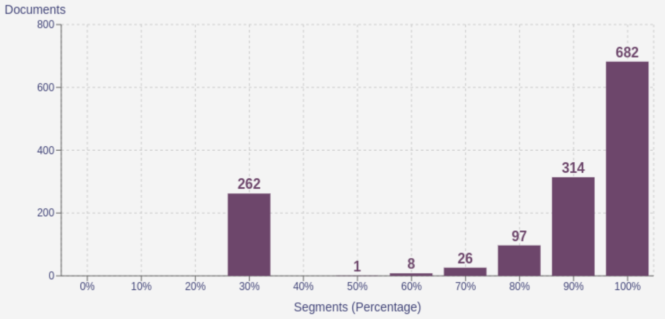
Language Distribution

Number of segments



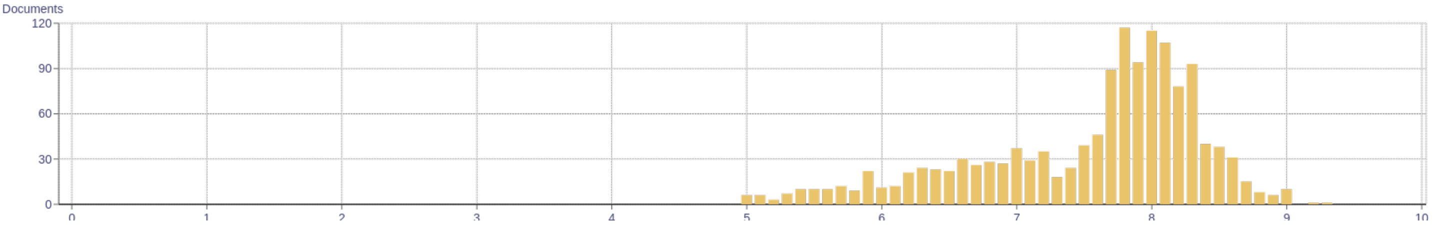
*Dyula (dyu) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Dyula (dyu) inside documents



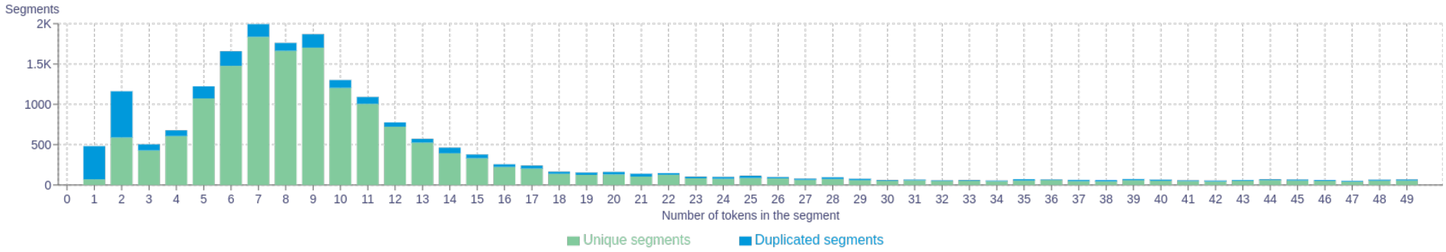
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.4K documents)

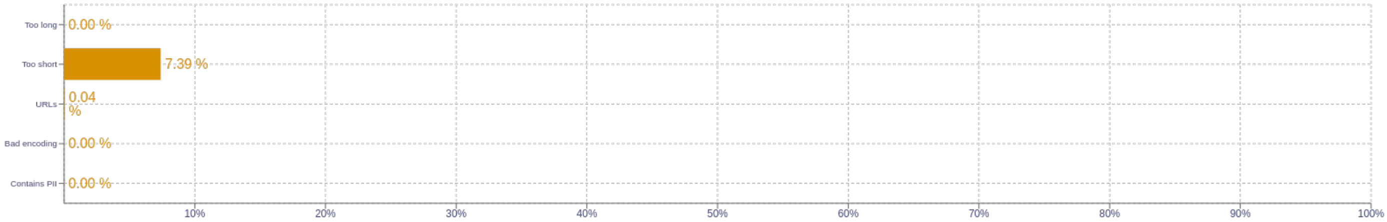


Segment length distribution by token

<= 49 tokens = **16K** segments | **2.8K** duplicates
> 50 tokens = **5.4K** segments | **1K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	o | 29654 u | 26214 be | 24276 ko | 22739 k | 19305
2	tun be | 4211 tuma min | 2018 o tigi | 1861 cogo min | 1771 min be | 1713
3	be se k | 824 u ye ko | 802 ala ka kuma | 789 ala ka masaya | 482 aw ye ko | 477
4	fɔra ala ka kuma | 267 ala ka mɔgɔ wolomanin | 151 masaba aw ka ala | 148 masaba le ko ten | 112 minɛ ka taga n | 100
5	ala ka kuma na ko | 183 ala ka mɔgɔ wolomanin nin | 99 ne masaba le ko ten | 55 aw ye ko ni mɔgɔ | 52 o yɔrɔ la ka taga | 51

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Volumes - Tokens

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Type-Token Ratio

Lexical variety computed as *number or types (uniques)/number of tokens*, after removing punctuation (https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (https://github.com/mbanon/fastspell).

Distribution of segments by fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by average fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

Segment length distribution by token

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Segment noise distribution

Obtained with Bicleaner Hardrules (https://github.com/bitextor/bicleaner-hardrules/).

Frequent n-grams

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md, after removing n-grams starting or ending in a stopword. Stopwords from https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt