HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lus_Latn	9/18/2025	Mizo (lus)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
294,926	5,467,373	4,121,383 (75.38 %)	223M	993,375,120	955.77 MB

Top 10 domains

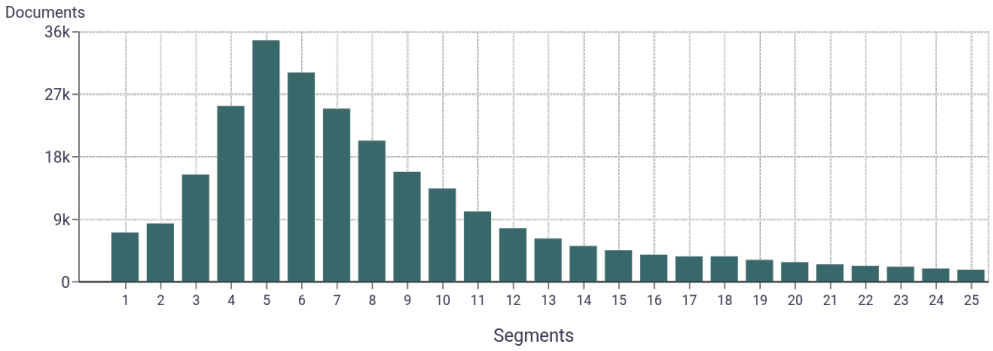
Domain	Docs	% of total
khampat.com	31K	10.60%
mizoram.gov.in	26K	8.93%
zomidaily.org	13K	4.28%
zonet.in	11K	3.58%
misual.com	9.8K	3.31%
zothlifim.com	7.9K	2.69%
blogspot.com	7.4K	2.51%
theaizawlpost.org	6.7K	2.28%
exploremizoram.com	4.8K	1.64%
thechinlandpost...	4.2K	1.44%

Top 10 TLDs

Domain	Docs	% of total
com	161K	54.63%
org	44K	14.75%
in	26K	8.96%
gov.in	26K	8.94%
info	16K	5.31%
net	7.1K	2.41%
co	2.5K	0.84%
co.in	1.9K	0.64%
com.au	1.5K	0.49%
xyz	1.3K	0.45%

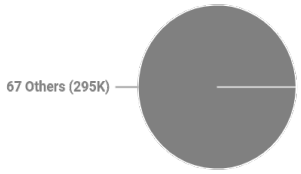
Documents size (in segments)

≤ 25 segments **87.29%** (257K documents)
> 25 segments **12.71%** (37K documents)



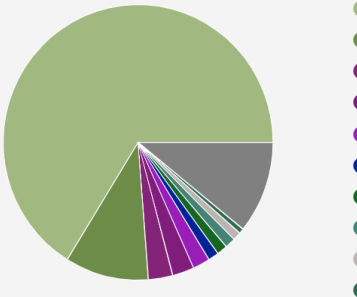
Document collections

CC = 91.10%
IA = 8.90%



Language Distribution

Number of segments in the Mizo (lus) corpus

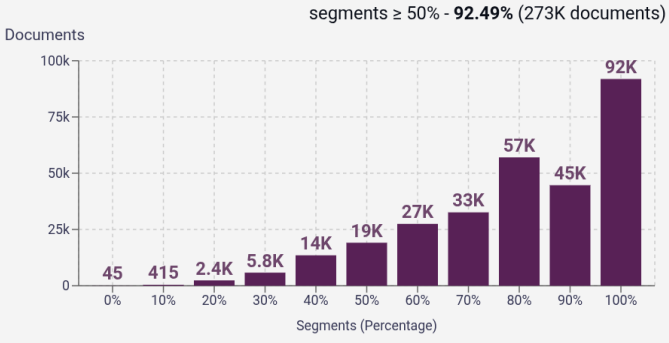


- English (en) - 3.6M **(66.2%)**
- Filipino (tl) - 547K **(10.0%)**
- Indonesian (id) - 160K **(2.9%)**
- Malay (ms) - 143K **(2.6%)**
- German (de) - 113K **(2.1%)**
- Spanish (es) - 70K **(1.3%)**
- Italian (it) - 69K **(1.3%)**
- French (fr) - 68K **(1.2%)**
- Welsh (cy) - 55K **(1.0%)**
- Slovenian (sl) - 32K **(0.6%)**
- 165 Others - 592K **(10.8%)**

*Mizo (lus) identification might be inaccurate because it is not supported by FastSpell

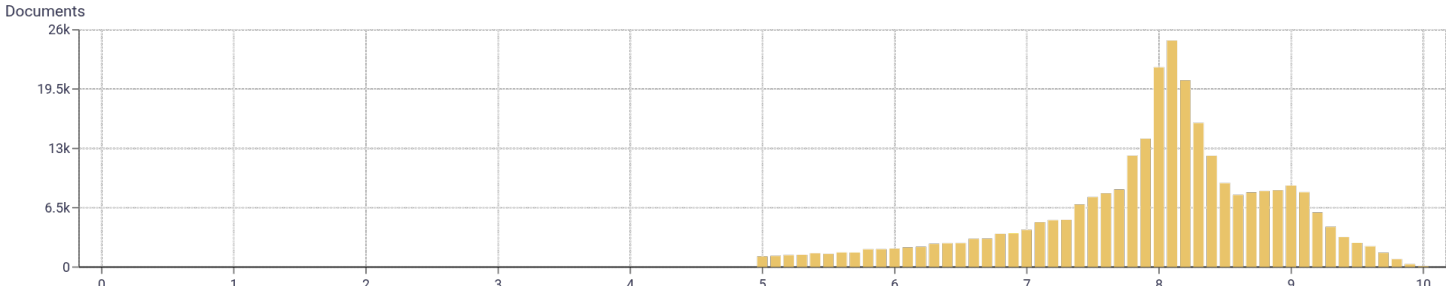
Percentage of segments in Mizo (lus) inside documents

segments < 50% - **7.51%** (22K documents)
segments ≥ 50% - **92.49%** (273K documents)



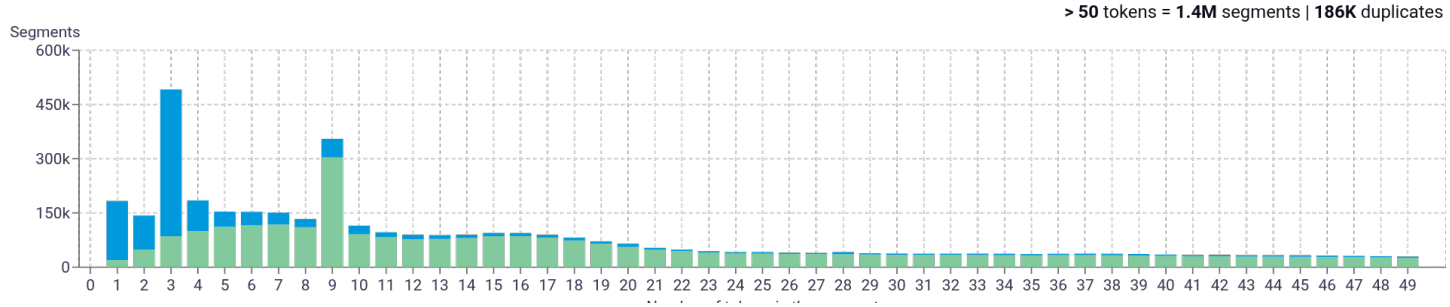
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (295K documents)

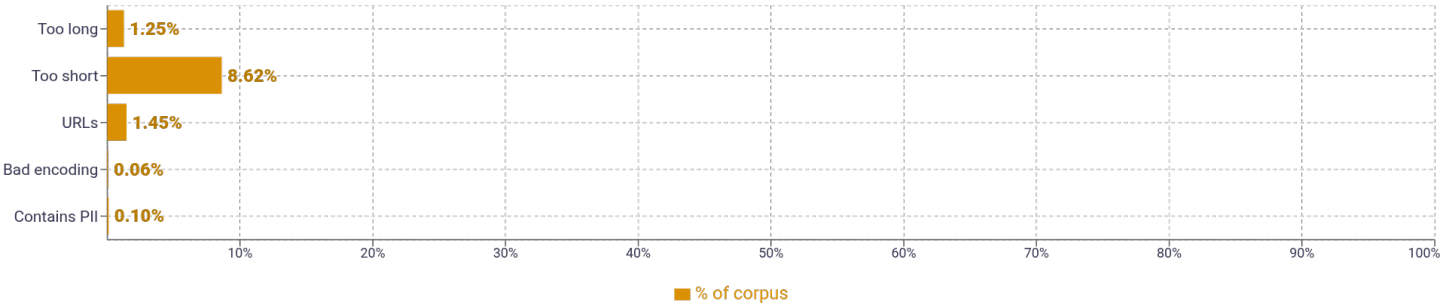


Segment length distribution by token

≤ 49 tokens = **4M** segments | **1.2M** duplicates
> 50 tokens = **1.4M** segments | **186K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	hi | 2,953,354 · leh | 2,061,313 · lo | 2,026,377 · chu | 1,678,203 · te | 1,436,885
2	this comment | 221,111 · report this | 221,029 · uh hi | 121,097 · te chu | 89,446 · ding hi | 89,110
3	report this comment | 221,025 · thu a sawi | 83,345 · post a comment | 44,705 · leave a reply | 25,274 · tiah a chim | 22,137
4	thu a sawi bawk | 19,851 · pakhat nih a chim | 9,338 · lawmawm a tih thu | 8,365 · tur thu a sawi | 7,035 · tur a ni lo | 5,196
5	chu lawmawm a tih thu | 7,416 · post chungchanga i ngaihdan lo | 4,825 · ngaihdan lo sawi ve rawh | 4,825 · i ngaihdan lo sawi ve | 4,825 · chungchanga i ngaihdan lo sawi | 4,825

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				