

General overview

Corpus	Date	Language
hplt-v3-mal_Mlym	9/18/2025	Malayalam

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
8,156,875	90,041,715	61,082,298 (67.84 %)	2.4B	18,998,254,672	47.81 GB

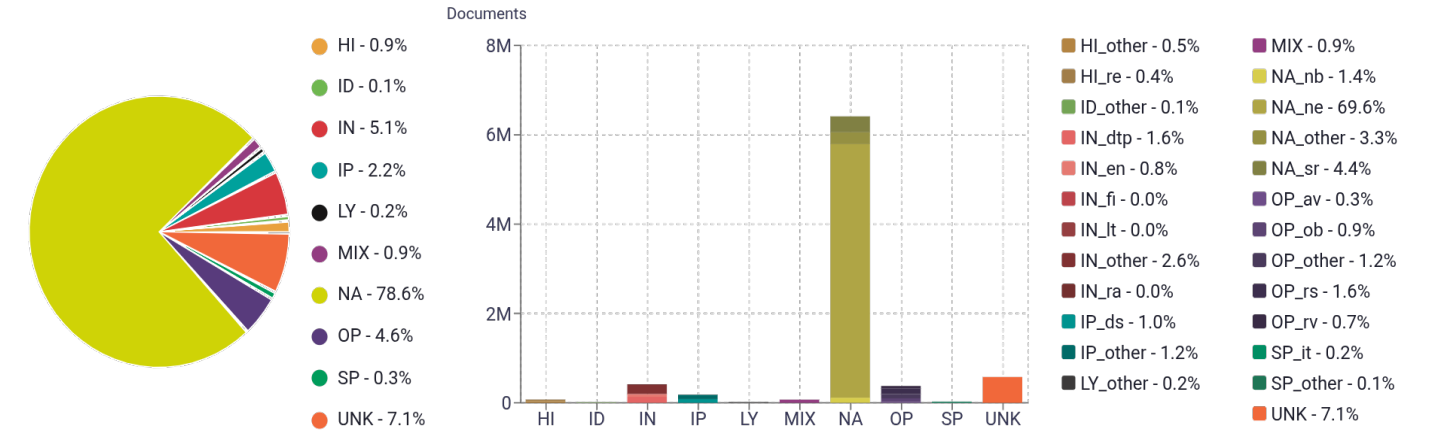
Top 10 domains

Domain	Docs	% of total
mathrubhumi.com	277K	3.39%
asianetnews.com	212K	2.60%
deepika.com	204K	2.50%
madhyamam.com	204K	2.50%
manoramaonline.com	181K	2.22%
blogspot.com	164K	2.02%
news18.com	129K	1.59%
twentyfournews.com	126K	1.54%
filmibeat.com	110K	1.35%
deshabhimani.com	99K	1.22%

Top 10 TLDs

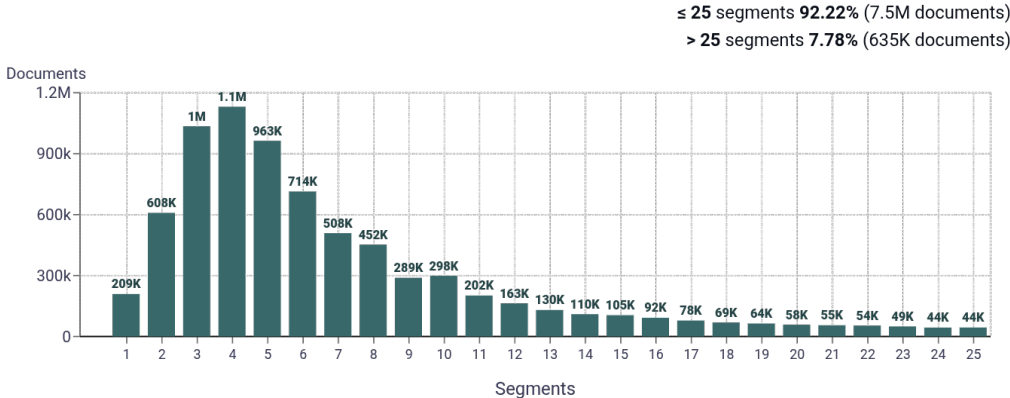
Domain	Docs	% of total
com	6.7M	82.04%
in	962K	11.79%
org	166K	2.03%
net	49K	0.60%
news	30K	0.37%
gov.in	30K	0.36%
online	22K	0.27%
co.uk	22K	0.27%
tv	22K	0.27%
live	20K	0.25%

Register labels

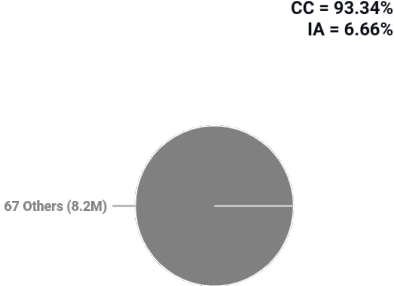


MT:2.4% | 198K Documents

Documents size (in segments) ⓘ

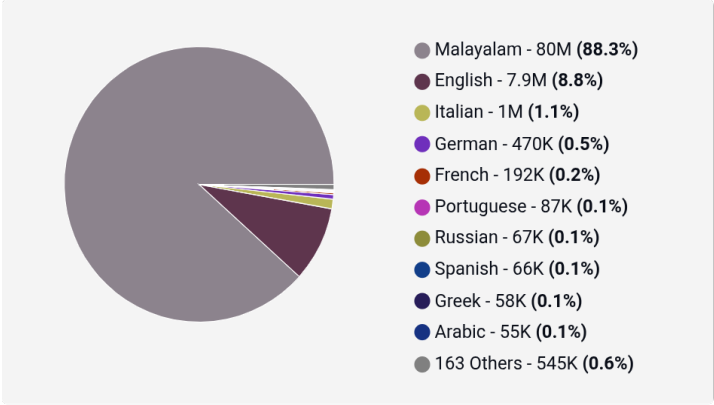


Document collections

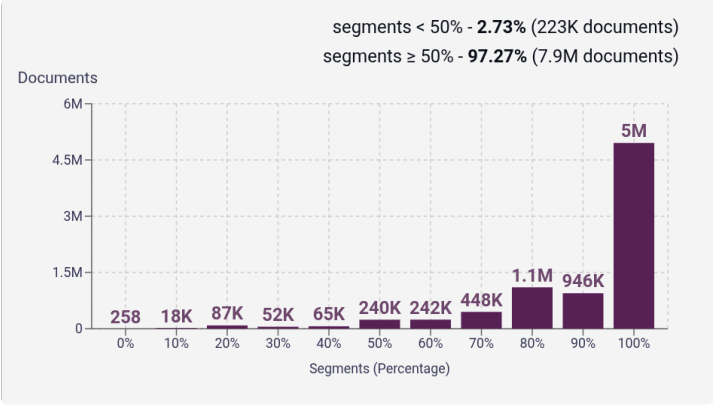


Language Distribution

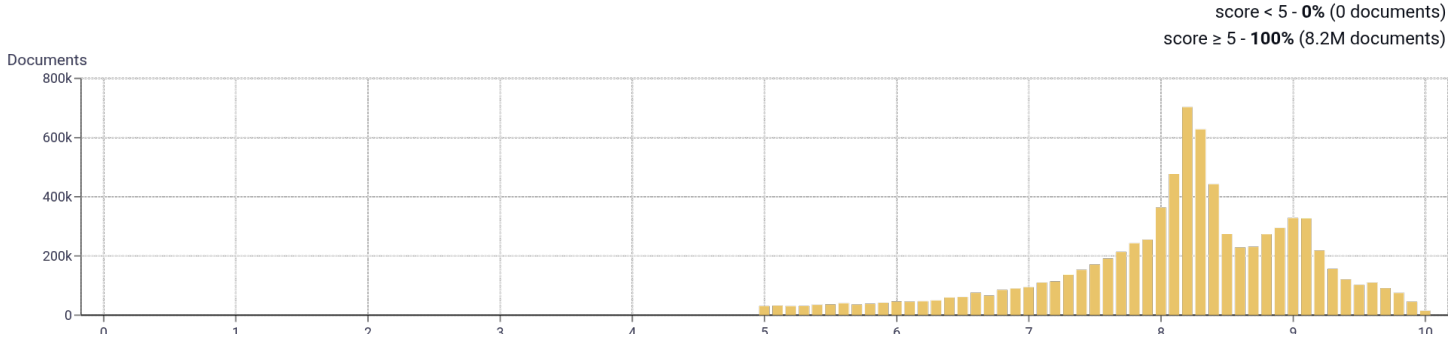
Number of segments in the Malayalam corpus



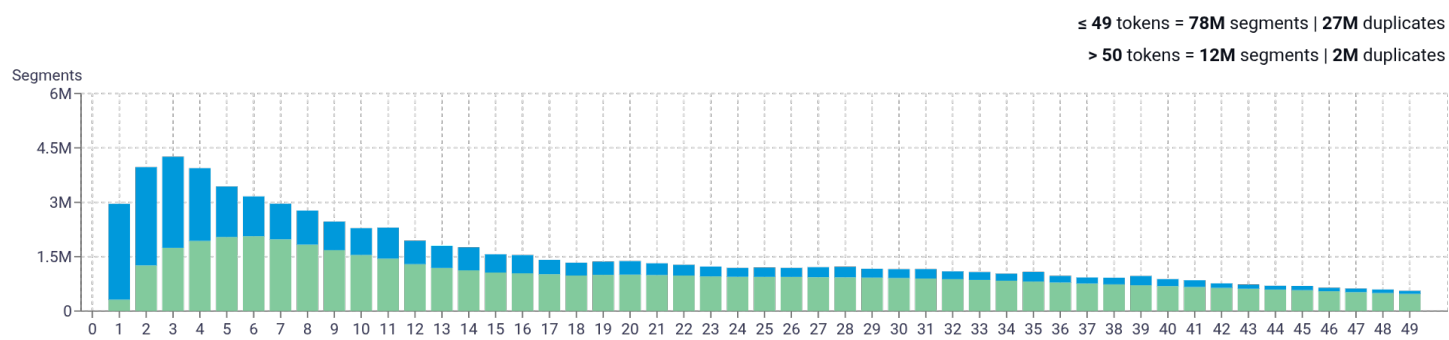
Percentage of segments in Malayalam inside documents



Distribution of documents by document score

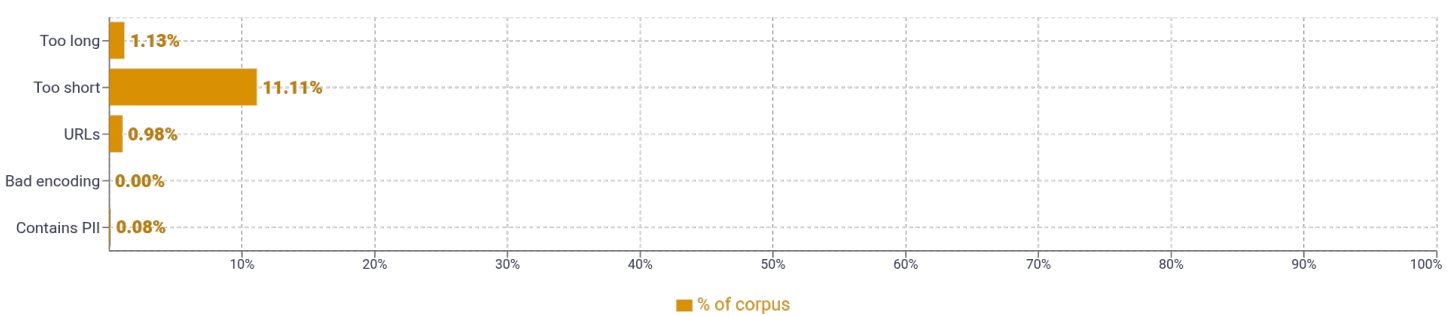


Segment length distribution by token



≤ 49 tokens = 78M segments | 27M duplicates
> 50 tokens = 12M segments | 2M duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	പറഞ്ഞു 4,798,804പി 4,234,696കെ 3,793,729സി 3,309,131എം 3,166,627	
2	read more 534,121കഴിഞ്ഞ ദിവസം 415,026കോടി രൂപ 283,433ലക്ഷം രൂപ 278,701a comment 247,154	
3	post a comment 207,046about this post 115,479post your comments 114,583സൈബർ നിയമപ്രകാരം ശിക്ഷാർഹമാണ് 112,763അശ്ലീലവും അസഭ്യവും നിയമവിരുദ്ധവും 109,163	
4	ഇത്തരം അഭിപ്രായങ്ങള് സൈബർ നിയമപ്രകാരം 108,102അഭിപ്രായങ്ങള് സൈബർ നിയമപ്രകാരം ശിക്ഷാർഹമാണ് 108,058പ്രതികരിക്കുന്നവർ അശ്ലീലവും അസഭ്യവും നിയമവിരുദ്ധവും 107,723വാർത്തകളോടു പ്രതികരിക്കുന്നവർ അശ്ലീലവും അസഭ്യവും 107,721അശ്ലീലവും അസഭ്യവും നിയമവിരുദ്ധവും അപകീർത്തികരവും 107,720	
5	ഇത്തരം അഭിപ്രായങ്ങള് സൈബർ നിയമപ്രകാരം ശിക്ഷാർഹമാണ് 108,058വാർത്തകളോടു പ്രതികരിക്കുന്നവർ അശ്ലീലവും അസഭ്യവും നിയമവിരുദ്ധവും 107,721പ്രതികരിക്കുന്നവർ അശ്ലീലവും അസഭ്യവും നിയമവിരുദ്ധവും അപകീർത്തികരവും 107,715അസഭ്യവും നിയമവിരുദ്ധവും അപകീർത്തികരവും സൂര്യ വളർത്തുന്നതുമായ 103,705നിയമവിരുദ്ധവും അപകീർത്തികരവും സൂര്യ വളർത്തുന്നതുമായ പരാമർശങ്ങള് 103,703	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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	dt
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				