HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-tam_Taml	9/18/2025	Tamil

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,266,205	205,078,222	130,153,460 (63.47 %)	4.4B	32,535,155,530	80.85 GB

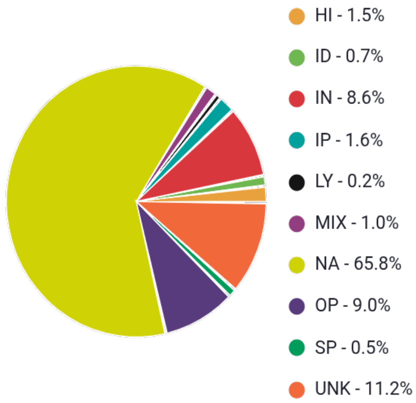
Top 10 domains

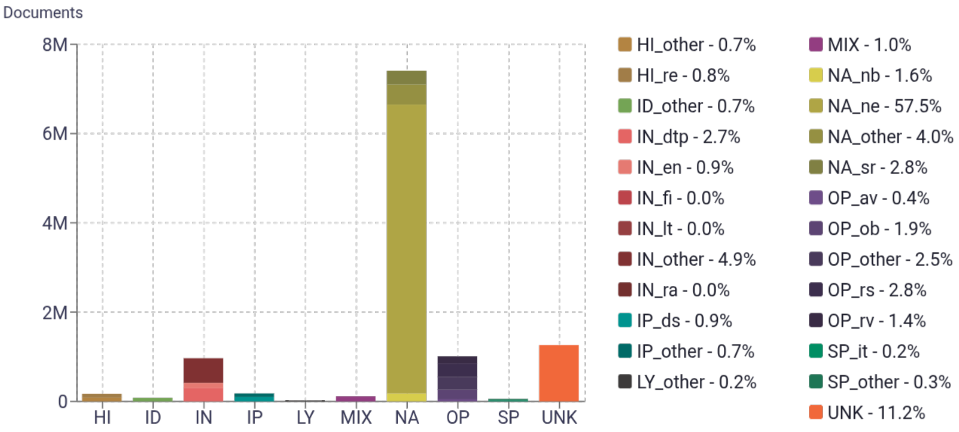
Domain	Docs	% of total
blogspot.com	546K	4.85%
dinakaran.com	350K	3.11%
dinamani.com	319K	2.83%
dinamalar.com	309K	2.74%
vikatan.com	239K	2.12%
hindutamil.in	197K	1.75%
maalaimalar.com	159K	1.41%
dailythanthi.com	152K	1.35%
asianetnews.com	143K	1.27%
news18.com	138K	1.22%

Top 10 TLDs

Domain	Docs	% of total
com	8.4M	74.88%
in	1M	9.07%
org	371K	3.29%
net	351K	3.12%
lk	332K	2.94%
ca	61K	0.55%
news	51K	0.45%
com.my	47K	0.42%
tv	46K	0.41%
ch	41K	0.37%

Register labels

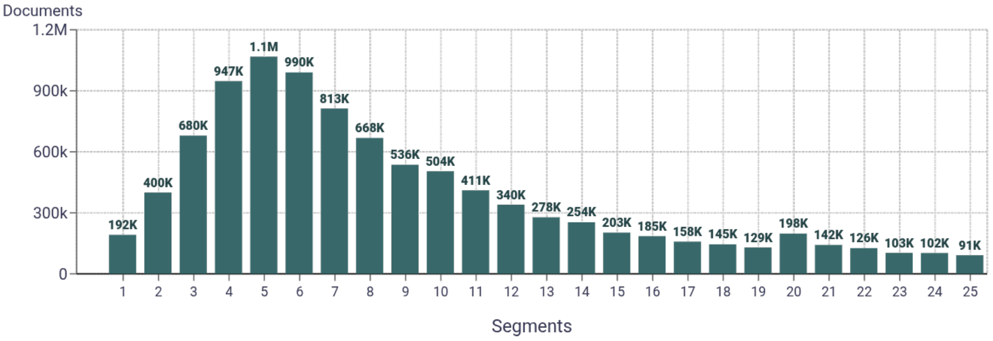




MT:3.8% | 425K Documents

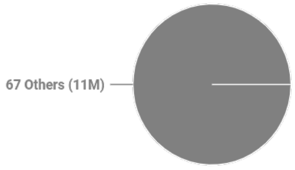
Documents size (in segments) ⓘ

≤ 25 segments **85.77%** (9.7M documents)
> 25 segments **14.23%** (1.6M documents)



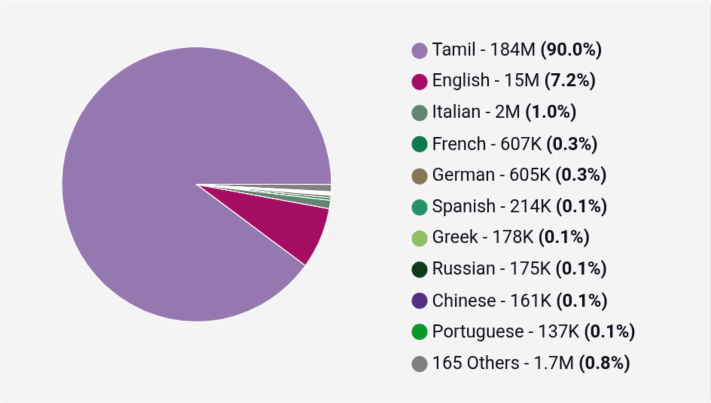
Document collections

CC = 91.26%
IA = 8.74%

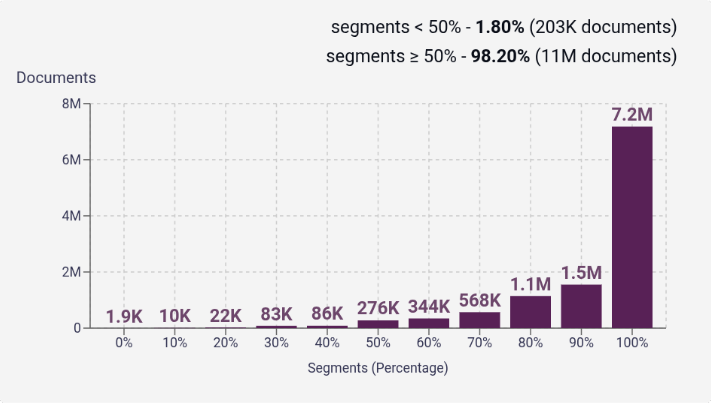


Language Distribution

Number of segments in the Tamil corpus



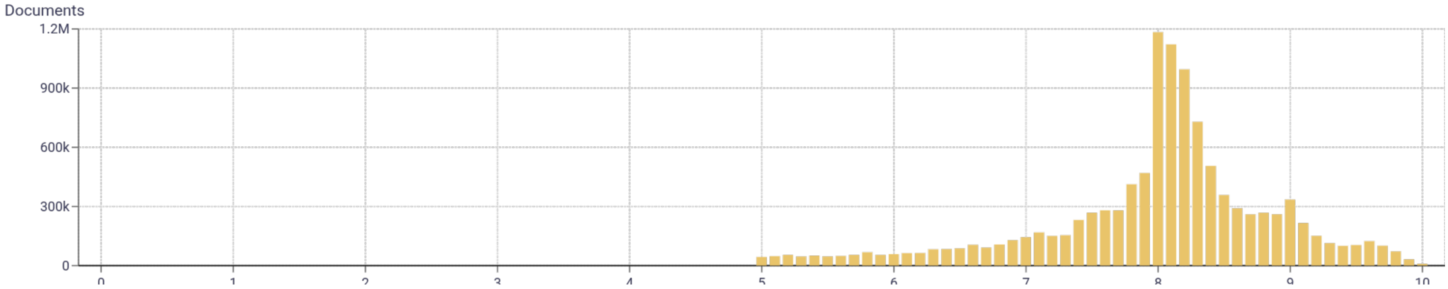
Percentage of segments in Tamil inside documents



Distribution of documents by document score

score < 5 - **0%** (0 documents)

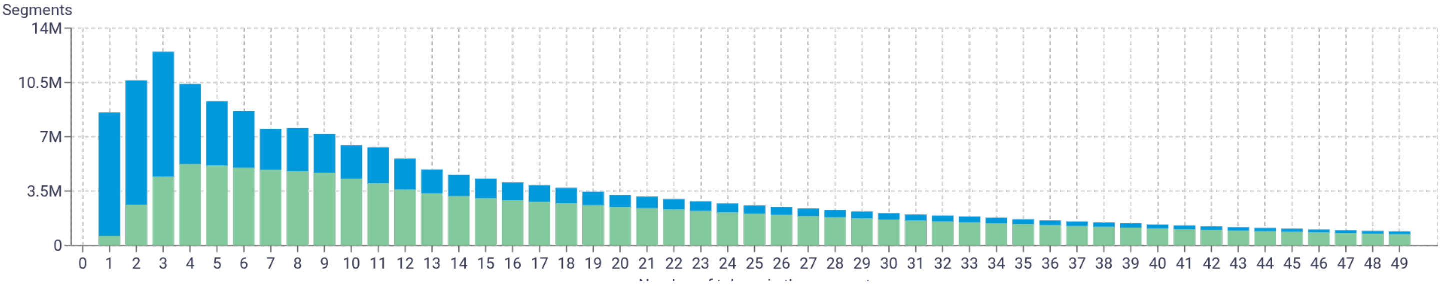
score ≥ 5 - **100%** (11M documents)



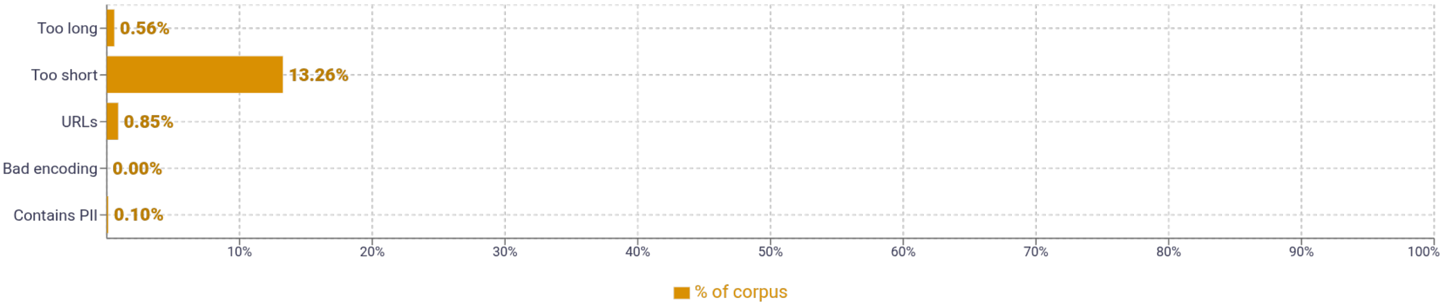
Segment length distribution by token

≤ **49** tokens = **185M** segments | **71M** duplicates

> **50** tokens = **20M** segments | **4.1M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ஆனால் | 7,592,065 செய்து | 5,283,014 இல்லை | 5,203,824 உங்கள் | 5,044,951 அரசு | 4,872,464
2	read more | 804,105 a comment | 799,956 post a | 683,261 மத்திய அரசு | 678,336 ஆம் ஆண்டு | 581,194
3	post a comment | 681,856 by ayyasamy ram | 190,947 in your inbox | 163,009 to receive our | 162,815 up to receive | 162,751
4	in your inbox every | 162,738 up to receive our | 162,736 your inbox every day | 162,735 to receive our newsletter | 162,735 receive our newsletter in | 162,735
5	up to receive our newsletter | 162,735 to receive our newsletter in | 162,735 receive our newsletter in your | 162,735 our newsletter in your inbox | 162,735 newsletter in your inbox every | 162,735

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				