

General overview

Corpus	Date	Language
hplt-v3-mni_Beng	9/18/2025	Manipuri

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7,573	189,316	121,758 (64.31 %)	5.7M	36,256,592	90.9 MB

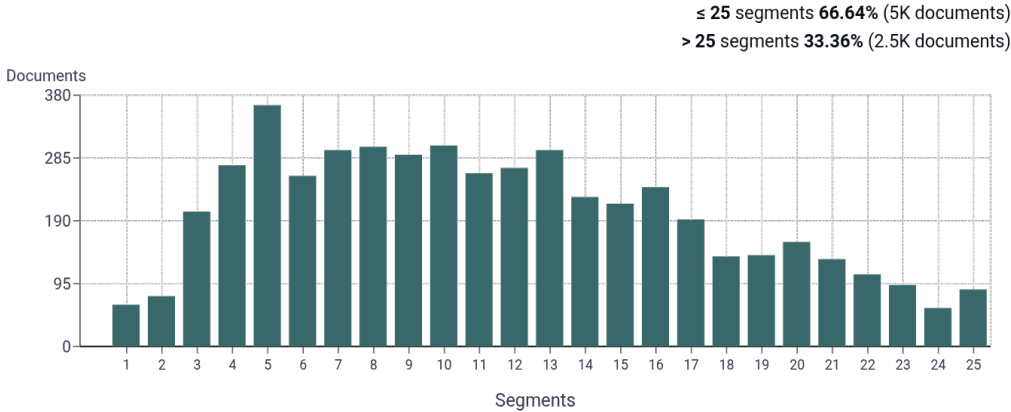
Top 10 domains

Domain	Docs	% of total
pib.gov.in	3.3K	43.22%
narendramodi.in	2.2K	28.91%
manipurimirror.com	973	12.85%
lakhipuronline.in	370	4.89%
vikaspedia.in	301	3.97%
pmindia.gov.in	45	0.59%
jw.org	38	0.50%
bible.is	37	0.49%
naharolghithouda...	33	0.44%
gotquestions.org	28	0.37%

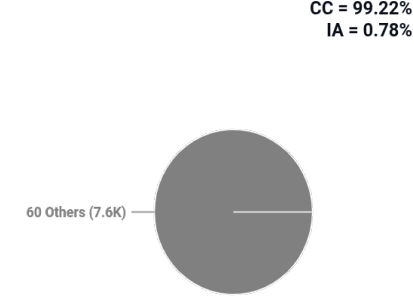
Top 10 TLDs

Domain	Docs	% of total
gov.in	3.3K	43.83%
in	2.9K	38.48%
com	1K	13.36%
nic.in	182	2.40%
org	103	1.36%
is	37	0.49%
ভারত	3	0.04%
org.in	1	0.01%
net	1	0.01%
com.bd	1	0.01%

Documents size (in segments) ⓘ

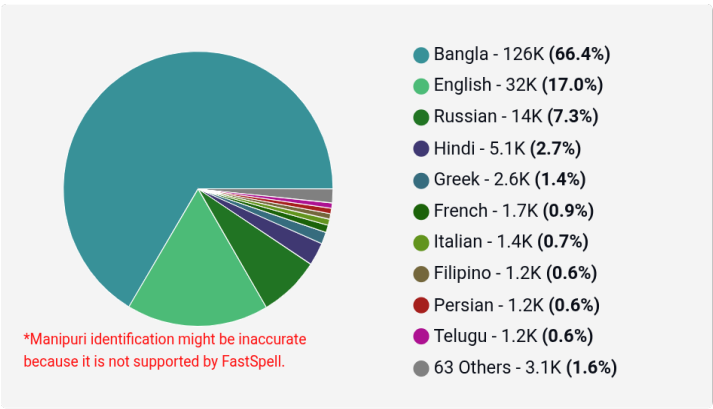


Document collections

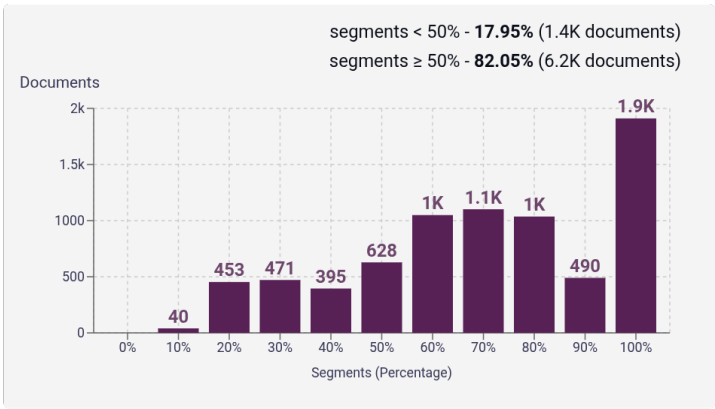


Language Distribution

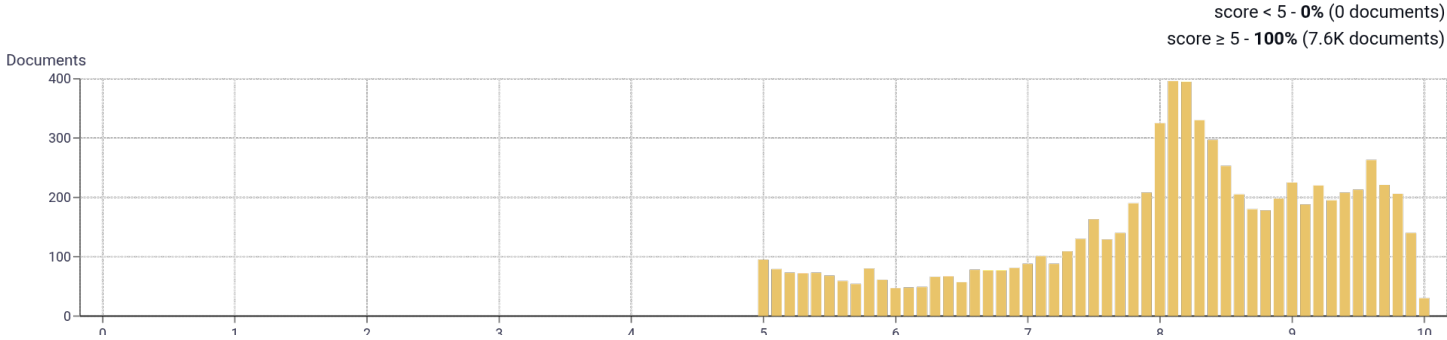
Number of segments in the Manipuri corpus



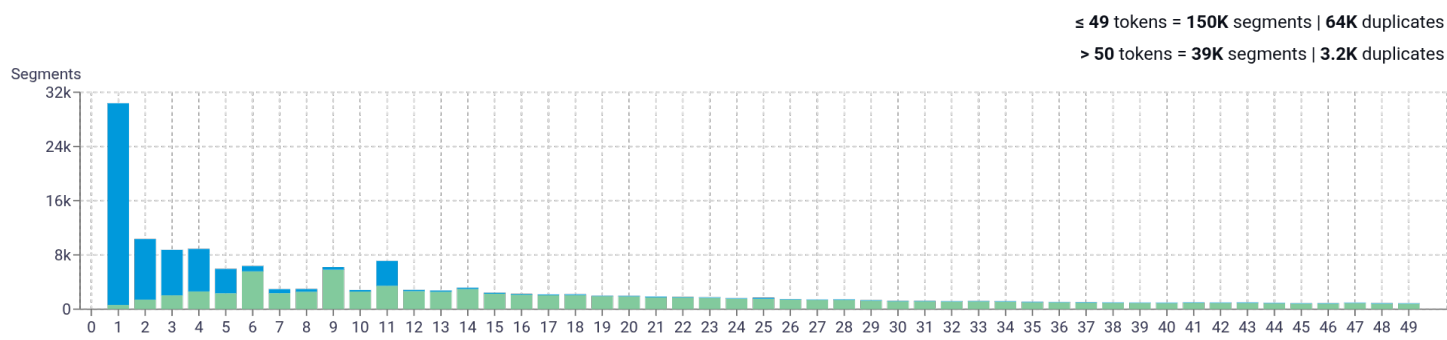
Percentage of segments in Manipuri inside documents



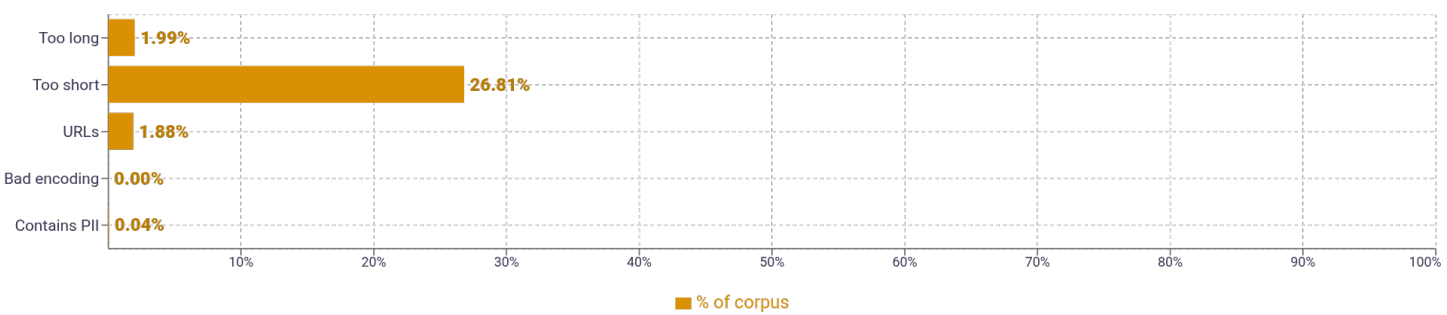
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	অমসুং 122,589 অসি 62,618 অমা 39,613 অসিদা 30,431 অসিনা 24,502	
2	প্রধান মন্ত্রীনা 12,901 লৈবাক অসিগী 5,563 লুপা করোর 5,211 মন্ত্রী লৈনবা 4,056 লৈবাক অসিদা 3,993	
3	by pib imphal 3,273 প্রধান মন্ত্রীনা হয়থি 3,248 শ্রী নরেন্দ্র মোদীনা 2,233 প্রধান মন্ত্রী শ্রী 1,999 হয়না প্রধান মন্ত্রীনা 1,697	
4	প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীনা 1,881 read this release in 1,203 প্রধান মন্ত্রীনা হয়থি মদুদি 763 হয়না মহারাধা মথা তাথি 519 হকশেল অমসুং যাইফ-মুখাল মন্ত্রালয় 488	
5	প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীনা 1,409 হৌথিবা পুং ২৪দা অনৌবা কেস 209 প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীগী 189 প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদী 188 ওসি প্রধান মন্ত্রী শ্রী নরেন্দ্র 153	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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	ntp
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				