HPLT Analytics report

HPLTAnalytics

General overview

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

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

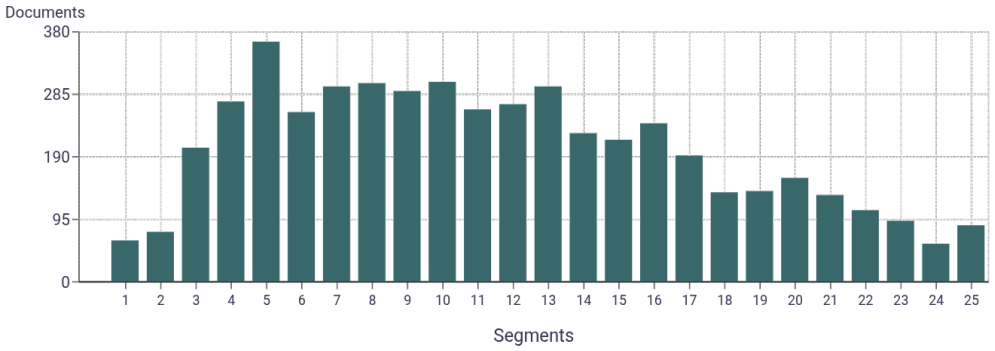
Domain	Docs	% of total
pib.gov.in	3.3K	43.22%
narendramodi.in	2.2K	28.91%
manipurimirror.com	973	12.85%
lakhipuronline.in	370	4.89%
vikaspedia.in	301	3.97%
pmindia.gov.in	45	0.59%
jw.org	38	0.50%
bible.is	37	0.49%
naharolgithouda...	33	0.44%
gotquestions.org	28	0.37%

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%

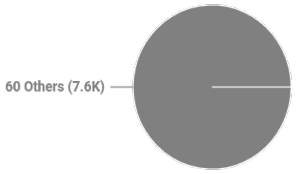
Documents size (in segments)

≤ 25 segments **66.64%** (5K documents)
> 25 segments **33.36%** (2.5K documents)



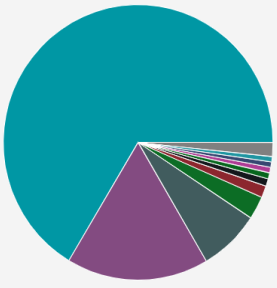
Document collections

CC = 99.22%
IA = 0.78%



Language Distribution

Number of segments in the Manipuri (mni) corpus

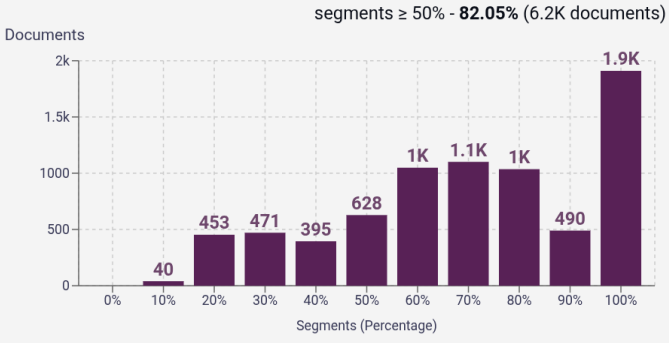


- Bangla (bn) - 126K **(66.4%)**
- English (en) - 32K **(17.0%)**
- Russian (ru) - 14K **(7.3%)**
- Hindi (hi) - 5.1K **(2.7%)**
- Greek (el) - 2.6K **(1.4%)**
- French (fr) - 1.7K **(0.9%)**
- Italian (it) - 1.4K **(0.7%)**
- Filipino (tl) - 1.2K **(0.6%)**
- Persian (fa) - 1.2K **(0.6%)**
- Telugu (te) - 1.2K **(0.6%)**
- 63 Others - 3.1K **(1.6%)**

*Manipuri (mni) identification might be inaccurate because it is not supported by FastSpell

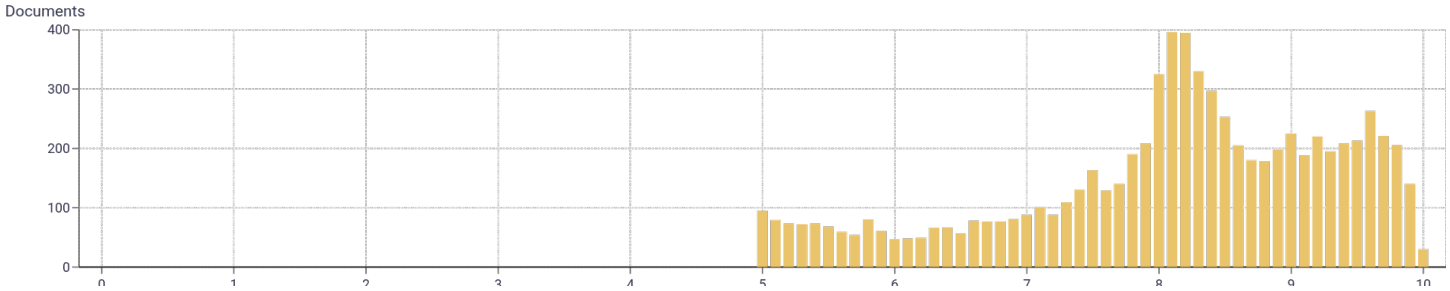
Percentage of segments in Manipuri (mni) inside documents

segments < 50% - **17.95%** (1.4K documents)
segments ≥ 50% - **82.05%** (6.2K documents)



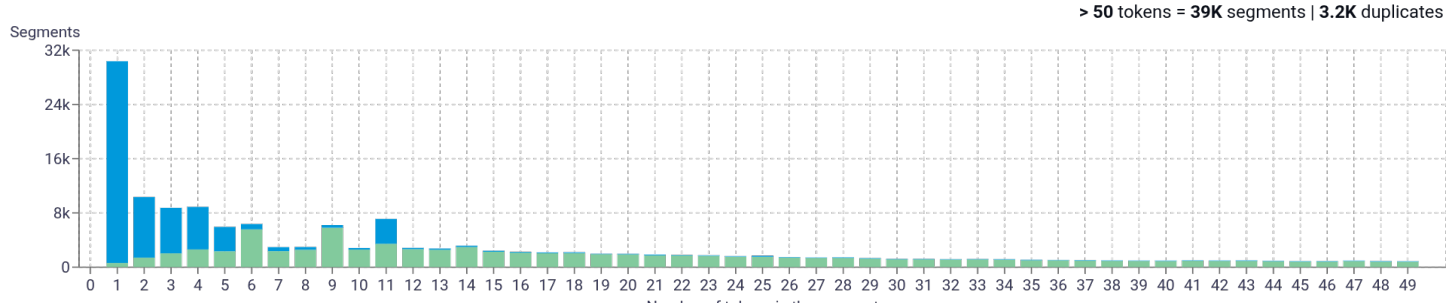
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (7.6K documents)

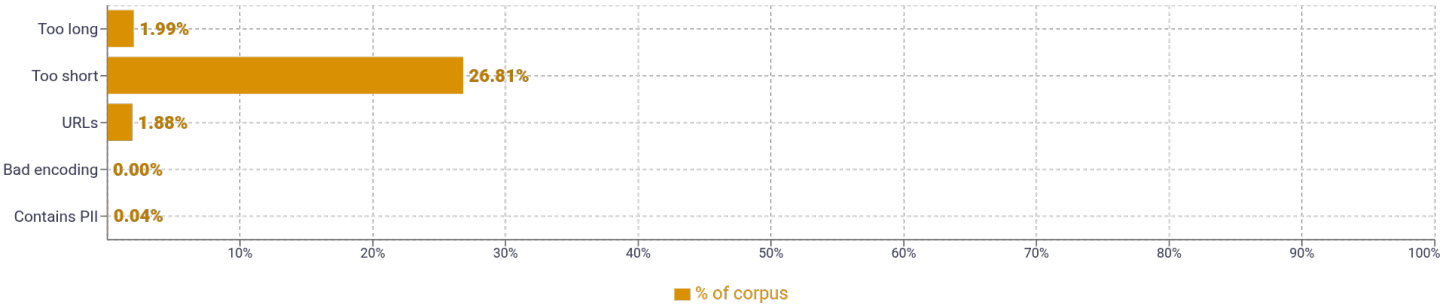


Segment length distribution by token

≤ 49 tokens = **150K** segments | **64K** duplicates
> 50 tokens = **39K** segments | **3.2K** duplicates



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 · মরী লৈনবা | 4,056 · লৈবাক অসিদা | 3,993
3	by pib imphal | 3,273 · প্রধান মন্ত্রীনা হায়খি | 3,248 · শ্রী নরেন্দ্র মোদীনা | 2,233 · প্রধান মন্ত্রী শ্রী | 1,999 · হায়না প্রধান মন্ত্রীনা | 1,697
4	প্রধান মন্ত্রী শ্রী নরেন্দ্র | 1,881 · read this release in | 1,203 · প্রধান মন্ত্রীনা হায়খি মদুদি | 763 · হায়না মহাক্কা মখা তাখি | 519 · হকশেল অমসুং য়াইফ-য়ুম্বাল মন্ত্রালয় | 488
5	প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীনা | 1,409 · হৌখিবা পুং ২৪দা অনৌবা কেস | 209 · প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদীগী | 189 · প্রধান মন্ত্রী শ্রী নরেন্দ্র মোদী | 188 · ঙসি প্রধান মন্ত্রী শ্রী নরেন্দ্র | 153

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				