

General overview

Corpus	Date	Language
hplt-v3-mai_Deva	9/18/2025	Maithili (mai)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
28,873	865,007	628,973 (72.71 %)	25M	115,987,171	282.34 MB

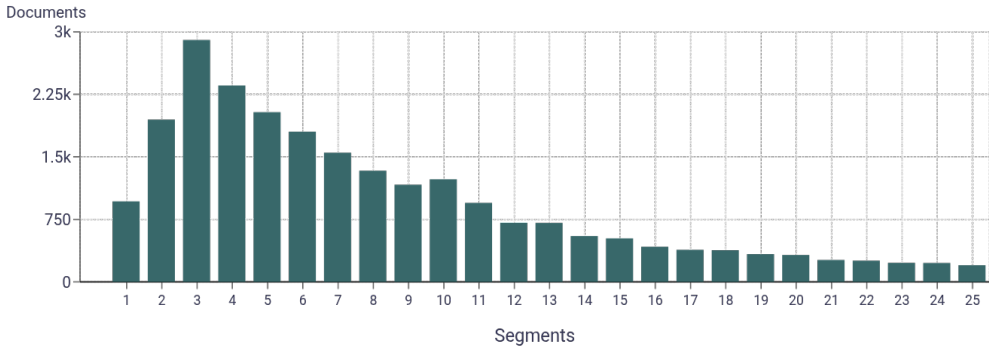
Top 10 domains

Domain	Docs	% of total
maithilijindaba...	5.1K	17.59%
blogspot.com	3.2K	11.18%
mithilamirror.com	2.7K	9.30%
esamaad.com	2.6K	9.05%
wikipedia.org	2.1K	7.34%
mithiladainik.in	1.4K	4.94%
mithimedia.in	1.1K	3.72%
hellomithila.com	1K	3.60%
videha.com	735	2.55%
ilovemithila.com	567	1.96%

Top 10 TLDs

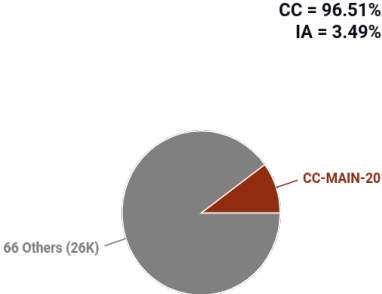
Domain	Docs	% of total
com	20K	69.78%
in	3.6K	12.54%
org	2.8K	9.60%
live	385	1.33%
pl	376	1.30%
net	301	1.04%
co.in	301	1.04%
de	228	0.79%
org.np	226	0.78%
com.np	213	0.74%

Documents size (in segments) ⓘ



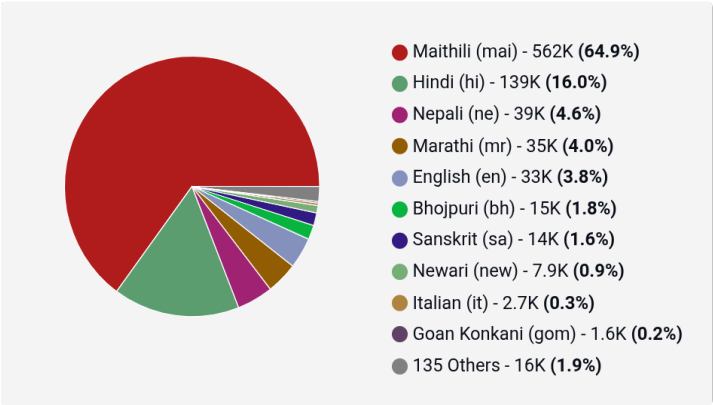
≤ 25 segments 82.24% (24K documents)  
> 25 segments 17.76% (5.1K documents)

Document collections

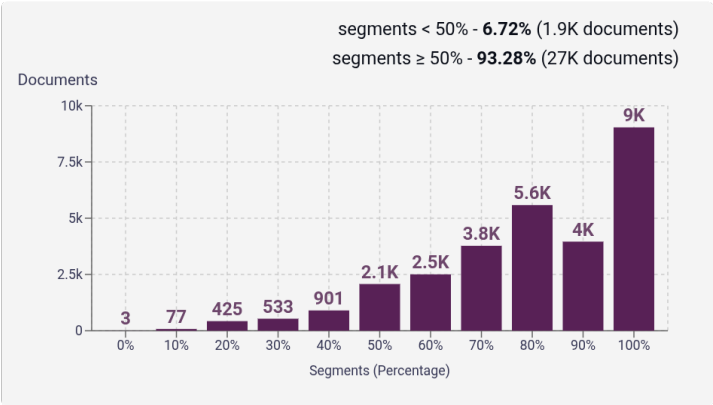


Language Distribution

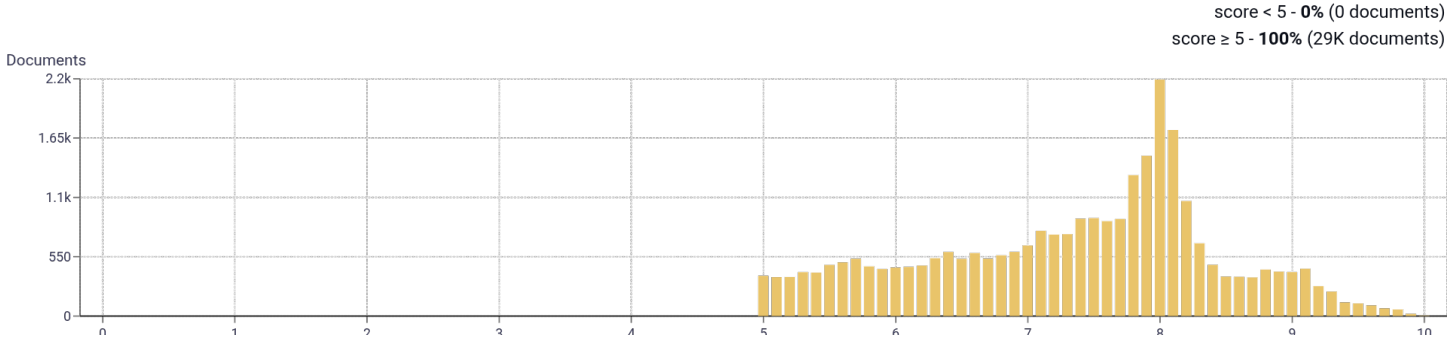
Number of segments in the Maithili (mai) corpus



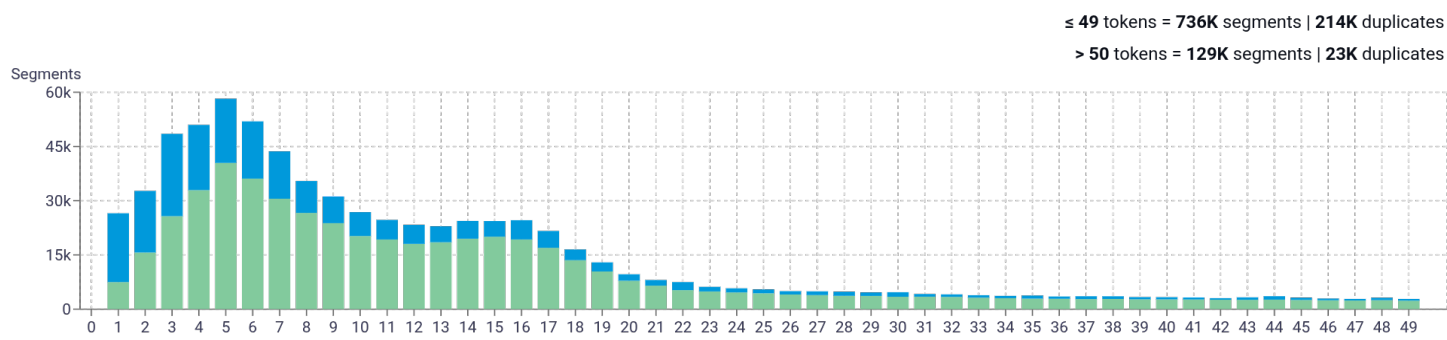
Percentage of segments in Maithili (mai) inside documents



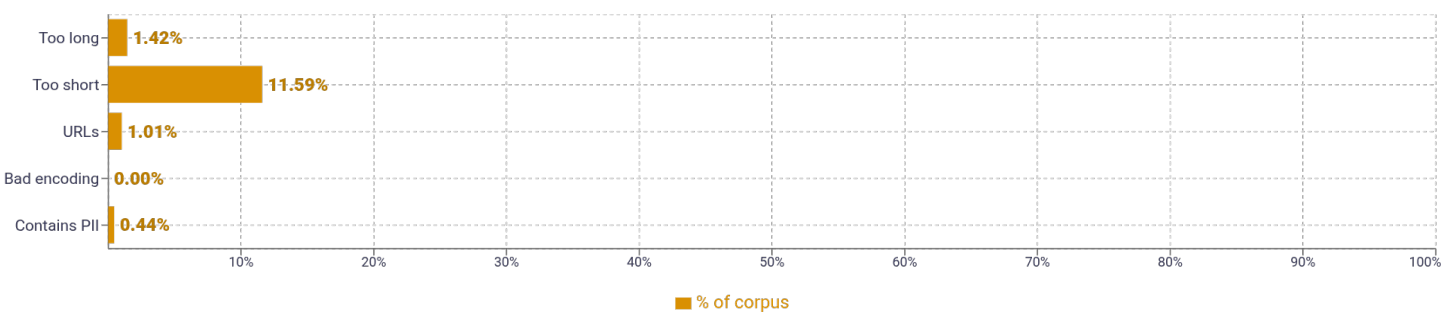
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	अछि   388,702   जे   197,745   सँ   170,731   ओ   119,262   ई   110,938	
2	रहल अछि   39,597   अछि जे   26,428   जाइत अछि   19,043   होइत अछि   14,056   सकैत अछि   12,728	
3	कएल गेल अछि   3,127   देल गेल अछि   3,081   post a comment   2,644   प्रवीण नारायण चौधरी   1,967   भऽ रहल अछि   1,892	
4	प्रथम मैथिली पाक्षिक ई   1,441   देल जा रहल अछि   1,275   सेहो पठा सकैत छी   1,101   com पर सेहो पठा   1,095   रचनात्मक सुझाव आ टीका   1,094	
5	com पर सेहो पठा सकैत   1,095   अपन रचनात्मक सुझाव आ टीका   1,094   प्रथम मैथिली पाक्षिक ई पत्रिका   892   मैथिलीक सभसँ लोकप्रिय आ सर्वग्राह्य   624 जालवृत्त पर प्रकाशित करबाक लेल   623	

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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>, <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	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				