

General overview

Corpus	Date	Language
hplt-v3-ben_Beng	9/17/2025	Bangla (bn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
25,557,260	358,969,558	227,888,609 (63.48 %)	11B	62,224,110,517	153.31 GB

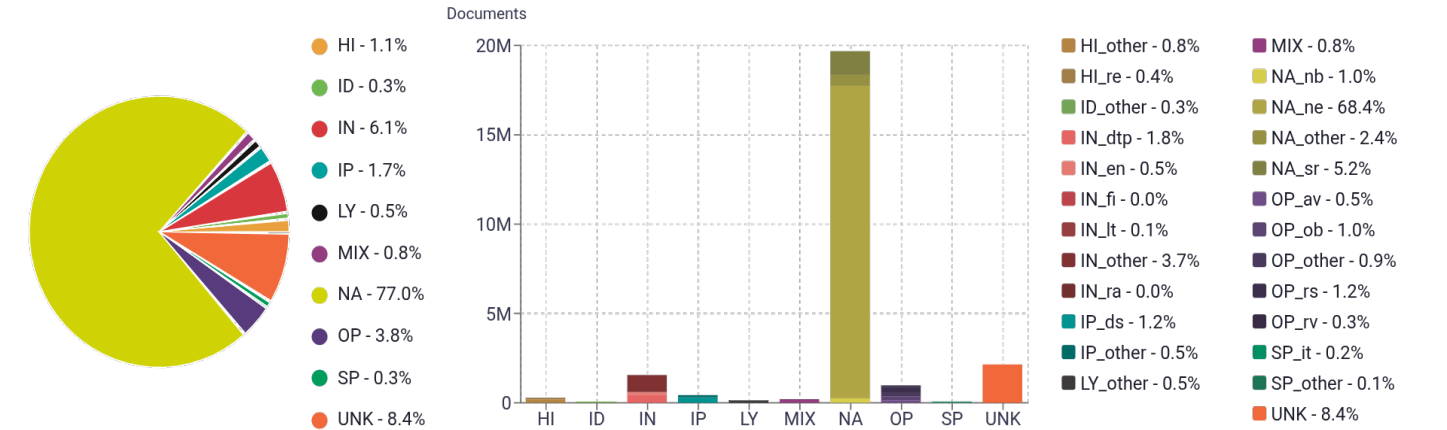
Top 10 domains

Domain	Docs	% of total
bdnews24.com	274K	1.07%
anandabazar.com	248K	0.97%
kalerkantho.com	204K	0.80%
banglanews24.com	194K	0.76%
prothomalo.com	185K	0.73%
bd-pratidin.com	171K	0.67%
hindustantimes.com	161K	0.63%
sangbadpratidin.in	159K	0.62%
news18.com	158K	0.62%
dailyjanakantha...	152K	0.60%

Top 10 TLDs

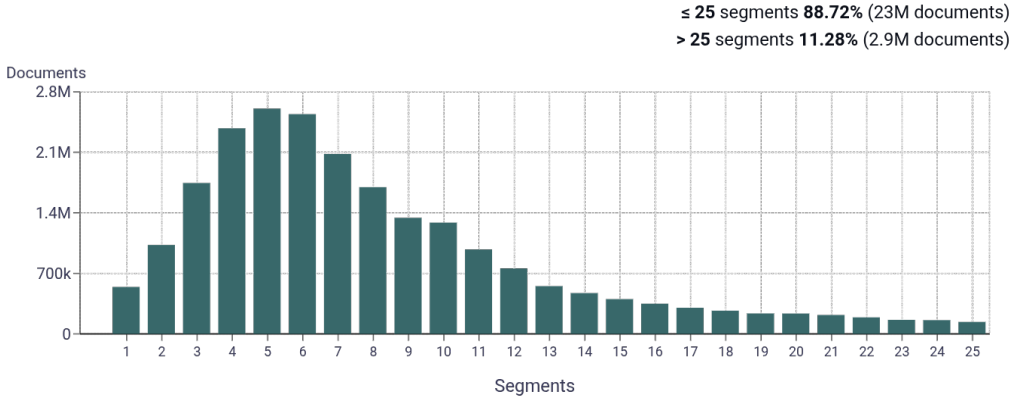
Domain	Docs	% of total
com	21M	81.13%
net	1.2M	4.76%
in	879K	3.44%
org	566K	2.22%
com.bd	526K	2.06%
tv	376K	1.47%
news	308K	1.20%
info	95K	0.37%
gov.bd	81K	0.32%
co	65K	0.26%

Register labels

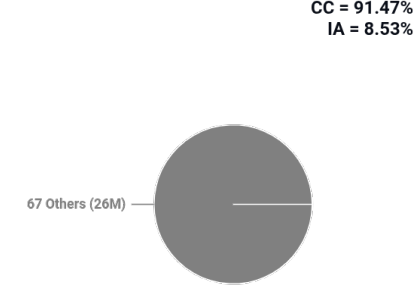


MT:4.7% | 1.2M Documents

Documents size (in segments) ⓘ

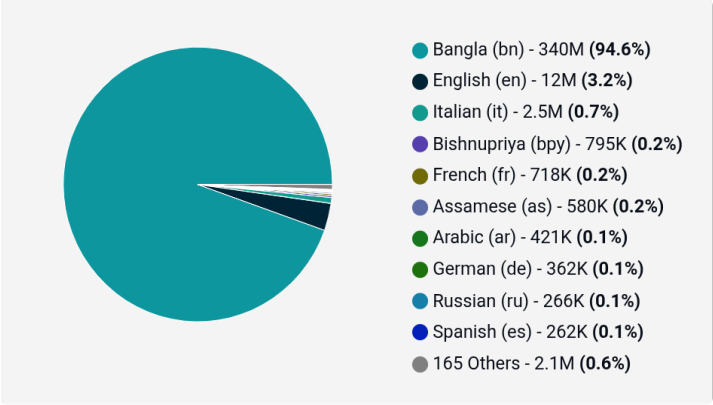


Document collections

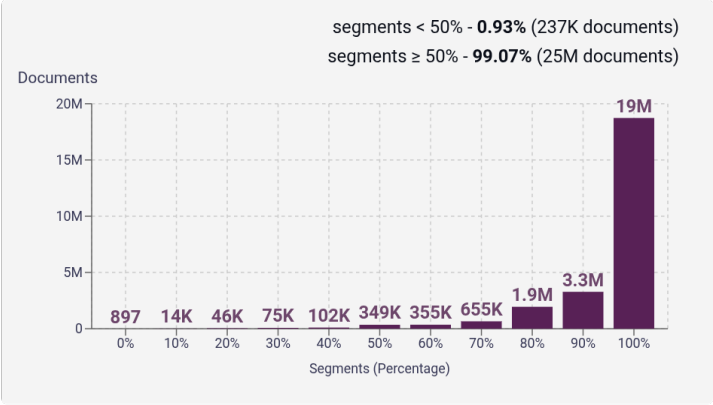


Language Distribution

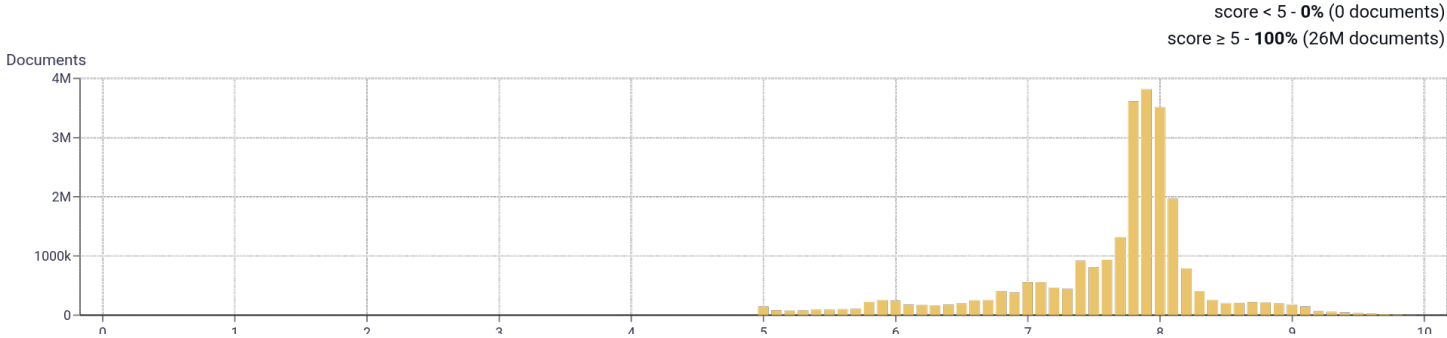
Number of segments in the Bangla (bn) corpus



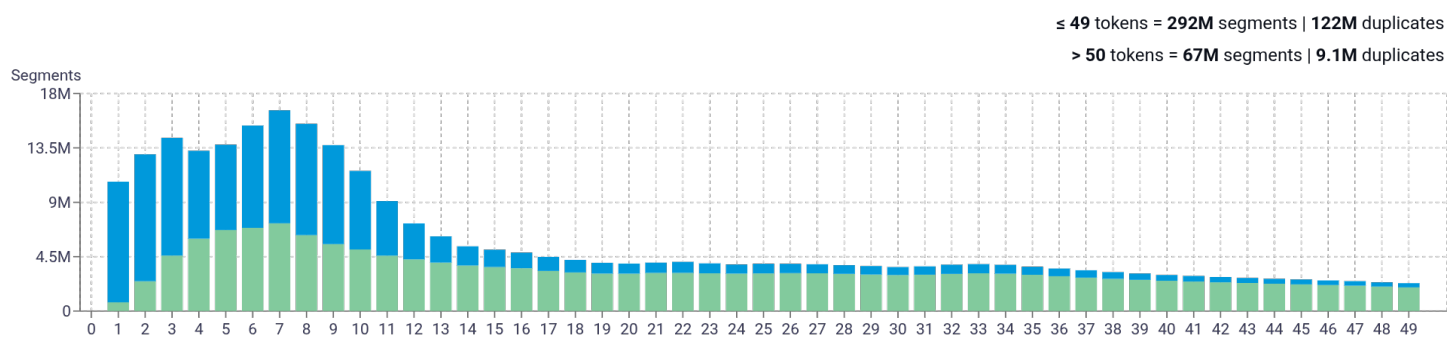
Percentage of segments in Bangla (bn) inside documents



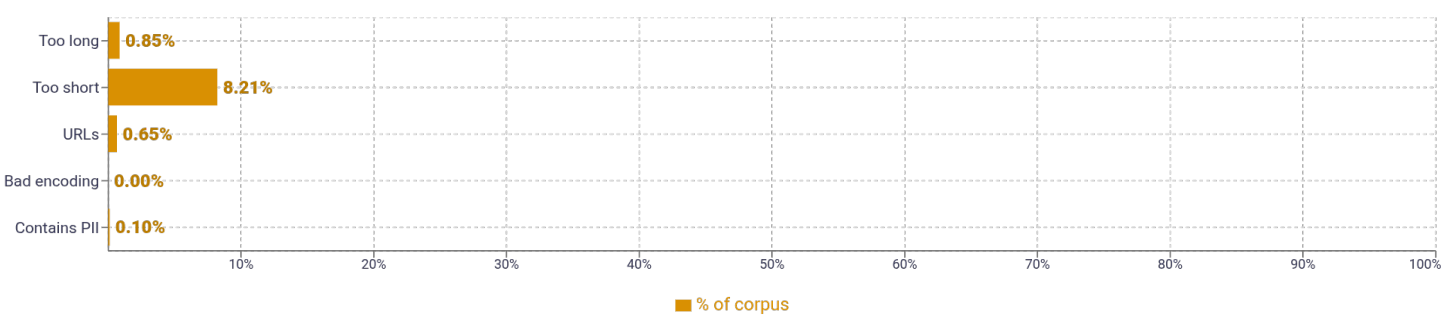
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS					
1	এক   26,033,410	সময়   19,395,091	সাথে   17,029,073	কথা   16,297,349	টাকা   13,352,979	
2	আওয়ামী লীগের   3,886,982	read more   2,631,156	আওয়ামী লীগ   2,274,452	শেখ হাসিনা   2,215,271	প্রধানমন্ত্রী শেখ   1,884,321	
3	প্রধানমন্ত্রী শেখ হাসিনা   1,145,059	বঙ্গবন্ধু শেখ মুজিবুর   851,513	অতিথি হিসেবে উপস্থিত   655,855	leave a reply   630,016	লীগের সাধারণ সম্পাদক   612,250	
4	আওয়ামী লীগের সাধারণ সম্পাদক   538,433	বঙ্গবন্ধু শেখ মুজিবুর রহমানের   460,203	প্রধান অতিথি হিসেবে উপস্থিত   396,591	জাতির পিতা বঙ্গবন্ধু শেখ   340,235		
	পিতা বঙ্গবন্ধু শেখ মুজিবুর   335,342					
5	জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর   329,353	পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমানের   208,192	মহাসচিব মির্জা ফখরুল ইসলাম আলমগীর   179,934			
	জাতির জনক বঙ্গবন্ধু শেখ মুজিবুর   179,608	সড়ক পরিবহন ও সেতুমন্ত্রী ওবায়দুল   125,801				

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 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>, <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				