

General overview

Corpus	Date	Language
hplt-v3-bho_Deva	9/17/2025	Bhojpuri

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
32,789	468,940	345,874 (73.76 %)	18M	80,088,951	192.73 MB

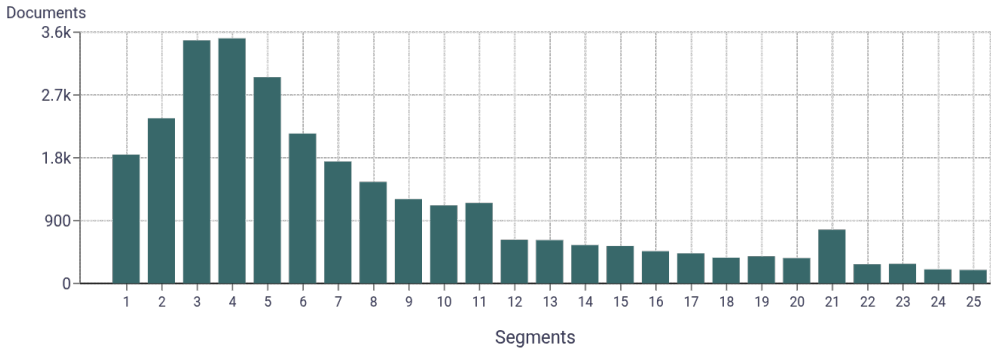
Top 10 domains

Domain	Docs	% of total
anjoria.com	11K	34.58%
khabarbhojpuri.com	2.2K	6.85%
wikipedia.org	2.1K	6.50%
news18.com	790	2.41%
jogira.com	655	2.00%
aapanbirgunj.com	593	1.81%
thebhojpuri.com	551	1.68%
bhojpurisahitya...	550	1.68%
ruralindiaonlin...	463	1.41%
bhojpuriyacinem...	419	1.28%

Top 10 TLDs

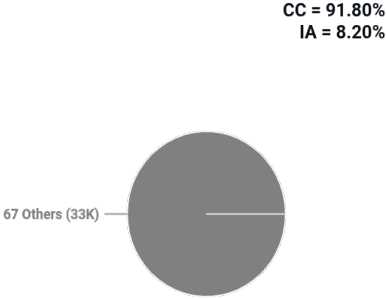
Domain	Docs	% of total
com	26K	79.47%
org	3.2K	9.90%
in	1.6K	4.80%
co.in	567	1.73%
page	245	0.75%
net	123	0.38%
mobi	123	0.38%
is	84	0.26%
mn	83	0.25%
ru	67	0.20%

Documents size (in segments) ⓘ



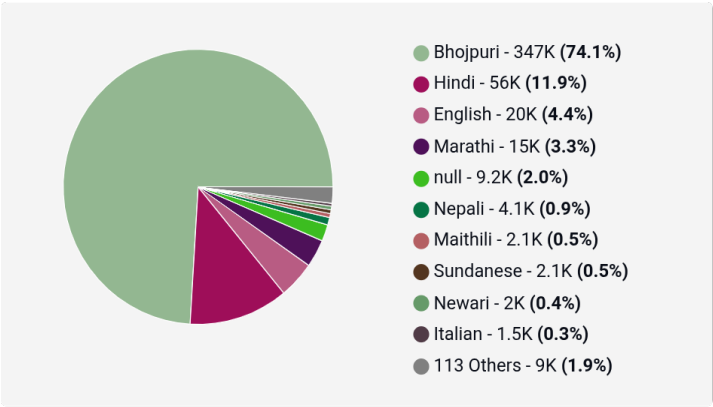
≤ 25 segments **88.6%** (29K documents)
> 25 segments **11.4%** (3.7K documents)

Document collections

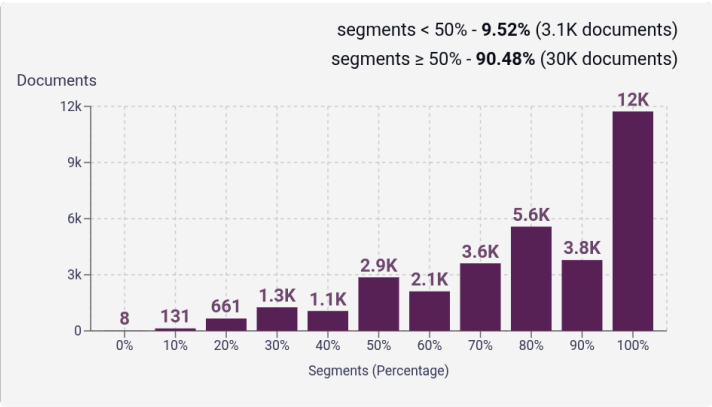


Language Distribution

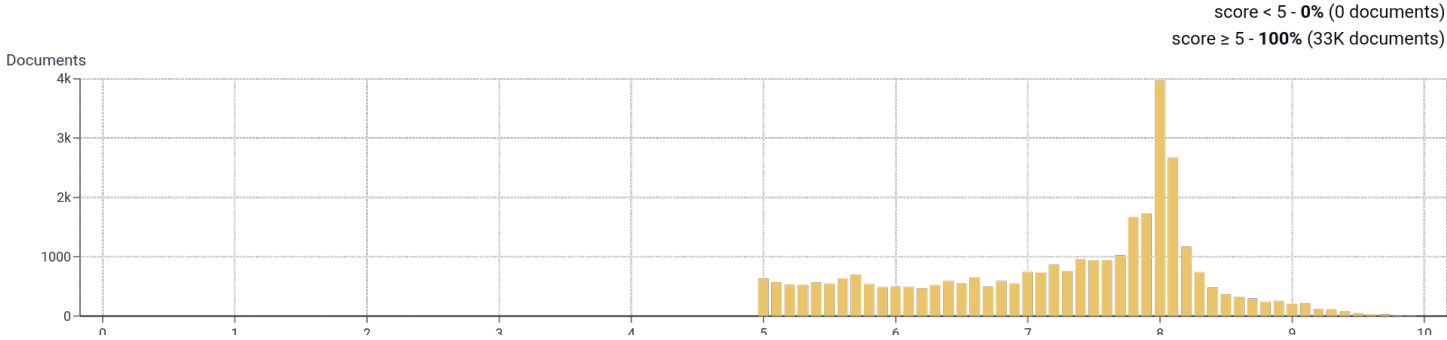
Number of segments in the Bhojpuri corpus



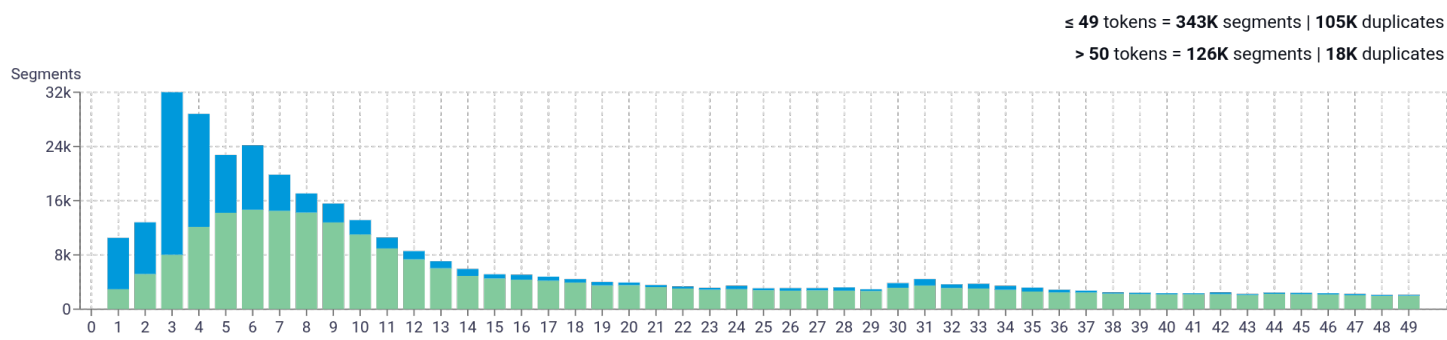
Percentage of segments in Bhojpuri inside documents



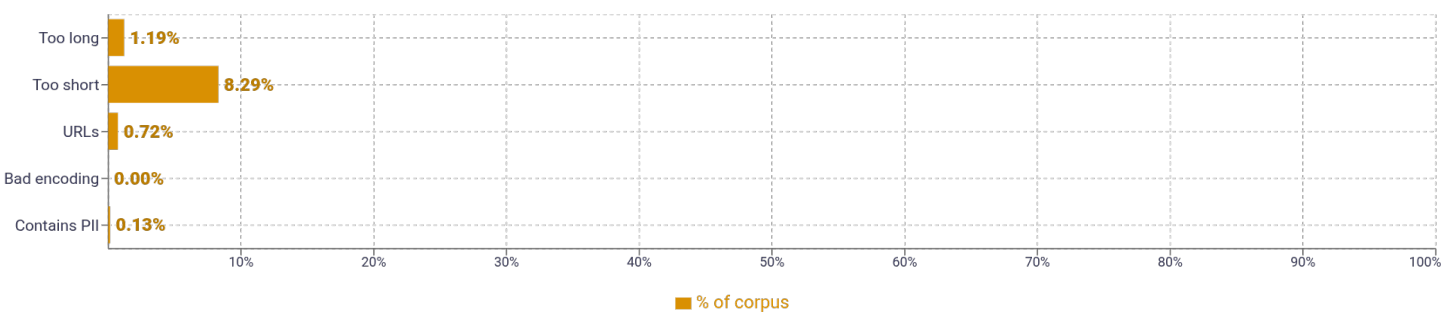
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	भोजपुरी 59,282 सिंह 54,840 कुमार 27,529 फिल्म 25,879 मे 22,333	
2	सेक्सी मूवी 4,319 भोजपुरी सिनेमा 3,859 वार्ड नंबर 3,135 मनोज तिवारी 3,084 लोक कवि 3,030	
3	रउरो कुछ कहीं... 1,572 सेक्सी मूवी हिंदी 1,139 भारतीय जनता पार्टी 1,134 हिंदी में सेक्सी 844 मुख्यमंत्री नीतीश कुमार 840	
4	सभकर थाह लगा गईनी 511 प्रणाम आ जय भोजपुरी 399 बिहार के मुख्यमंत्री नीतीश 396 सेक्सी फिल्म फुल एचडी 367 हिंदी में सेक्सी मूवी 363	
5	बिहार के मुख्यमंत्री नीतीश कुमार 379 comments posted to this article 215 सेक्सी मूवी हिंदी में फुल 209 साहित्य पढ़े जोगीरा के फेसबुक पेज 206 पढ़े जोगीरा के फेसबुक पेज 206	

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				