

General overview

Corpus	Date	Language
hplt-v3-awa_Deva	9/16/2025	Awadhi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
34,188	354,215	311,110 (87.83 %)	14M	64,703,194	153.44 MB

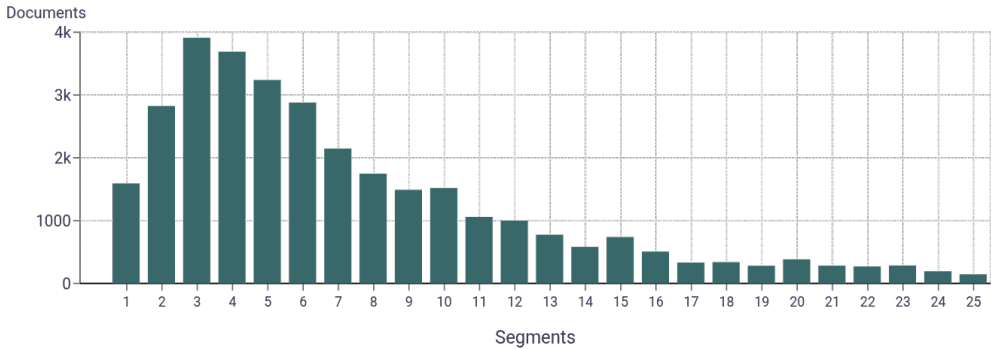
Top 10 domains

Domain	Docs	% of total
khabarlahariya.org	2.2K	6.50%
wikipedia.org	941	2.75%
newsbytesapp.com	692	2.02%
districtsinindi...	679	1.99%
biblegateway.com	653	1.91%
sportskeeda.com	506	1.48%
indiatimes.com	466	1.36%
jagran.com	439	1.28%
india.com	337	0.99%
khaskhabar.com	326	0.95%

Top 10 TLDs

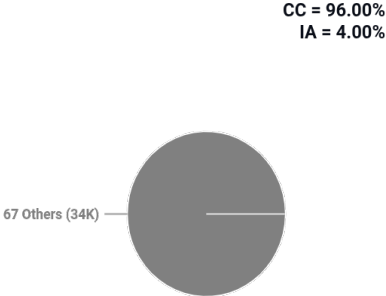
Domain	Docs	% of total
com	23K	67.68%
in	4.5K	13.27%
org	4.3K	12.71%
co.in	382	1.12%
net	337	0.99%
is	228	0.67%
news	227	0.66%
page	180	0.53%
nic.in	116	0.34%
co	93	0.27%

Documents size (in segments) ⓘ



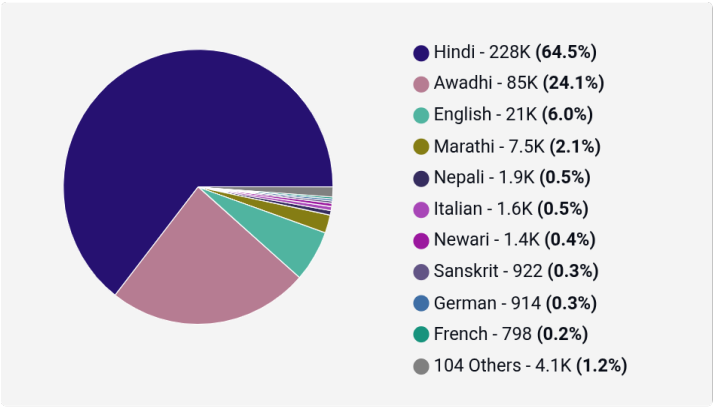
≤ 25 segments **94.21%** (32K documents)
> 25 segments **5.79%** (2K documents)

Document collections

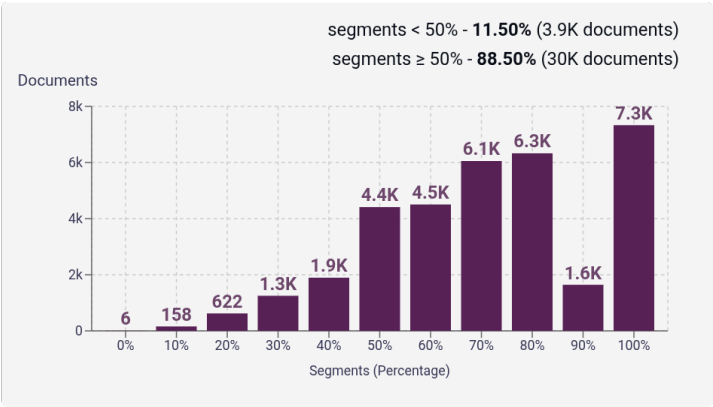


Language Distribution

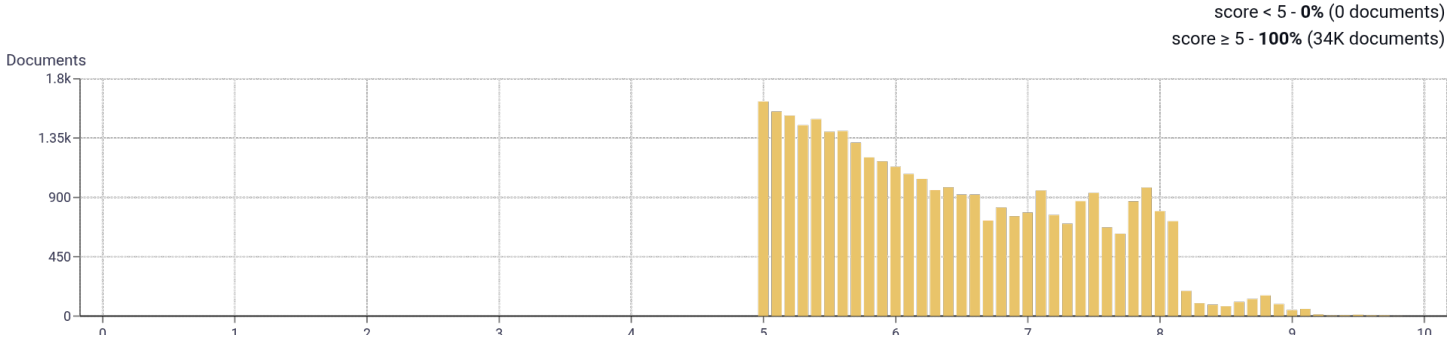
Number of segments in the Awadhi corpus



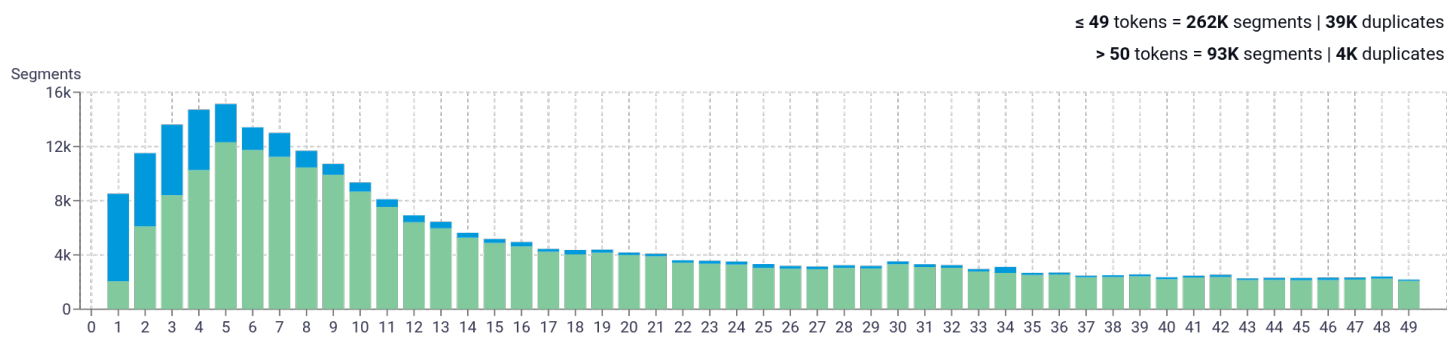
Percentage of segments in Awadhi inside documents



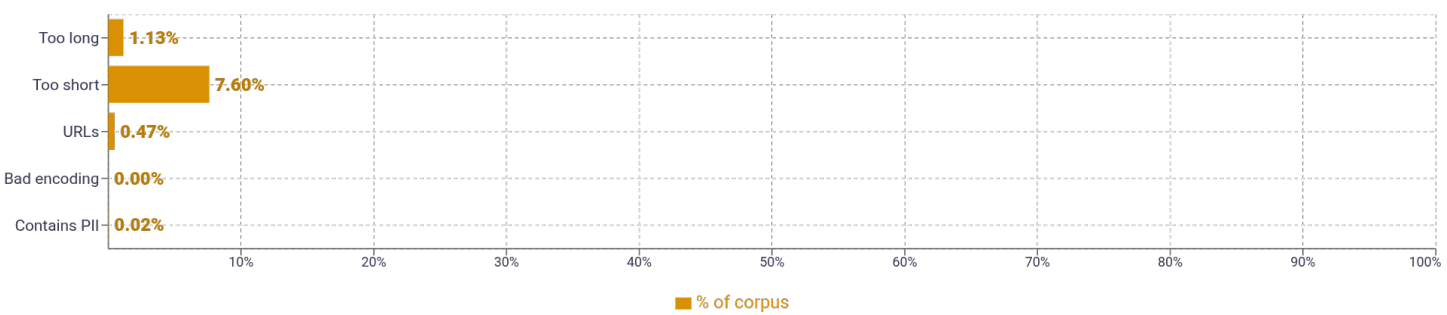
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>और 154,532</div> <div>ने 144,858</div> <div>भारत 72,231</div> <div>टेस्ट 61,629</div> <div>रन 57,368</div>	
2	<div>रन बनाए 15,978</div> <div>उत्तर प्रदेश 13,960</div> <div>रोहित शर्मा 12,254</div> <div>टेस्ट मैच 9,944</div> <div>भारत ने 9,476</div>	
3	<div>रोहित शर्मा ने 3,863</div> <div>भारतीय जनता पार्टी 3,803</div> <div>रन बनाए थे 3,641</div> <div>इंग्लैंड के खिलाफ 3,092</div> <div>मैचों की सीरीज 2,945</div>	
4	<div>रन की पारी खेली 1,811</div> <div>हार का सामना करना 1,695</div> <div>मैचों की टेस्ट सीरीज 1,641</div> <div>दक्षिण अफ्रीका के खिलाफ 1,586</div> <div>रनों की पारी खेली 1,439</div>	
5	<div>हार का सामना करना पड़ा 1,661</div> <div>विधान सभा चुनाव में इन्होंने 897</div> <div>रन की पारी खेली थी 717</div> <div>रनों की पारी खेली थी 693</div> <div>उत्तर प्रदेश विधान सभा चुनाव 676</div>	

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				