

General overview

Corpus	Date	Language
hplt-v3-kan_Knda	9/18/2025	Kannada

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,364,190	56,887,954	39,000,850 (68.56 %)	1.5B	9,953,831,723	24.57 GB

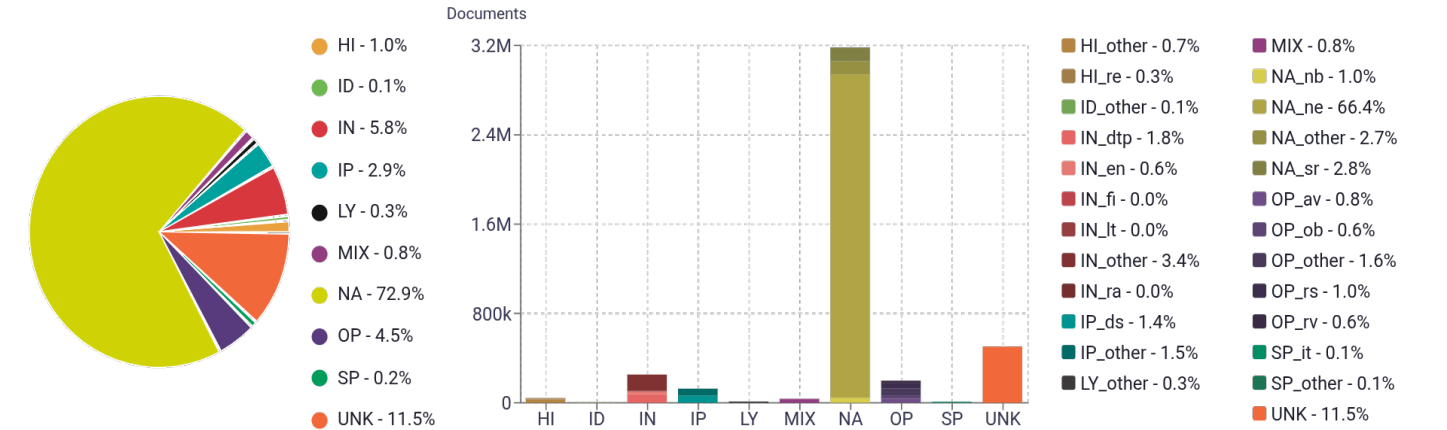
Top 10 domains

Domain	Docs	% of total
prajavani.net	373K	8.55%
udayavani.com	215K	4.93%
asianetnews.com	195K	4.46%
oneindia.com	118K	2.70%
news18.com	108K	2.48%
varthabharati.in	107K	2.44%
publiktv.in	90K	2.06%
filmibeat.com	90K	2.05%
vijaykarnataka.com	82K	1.87%
sanjevani.com	63K	1.43%

Top 10 TLDs

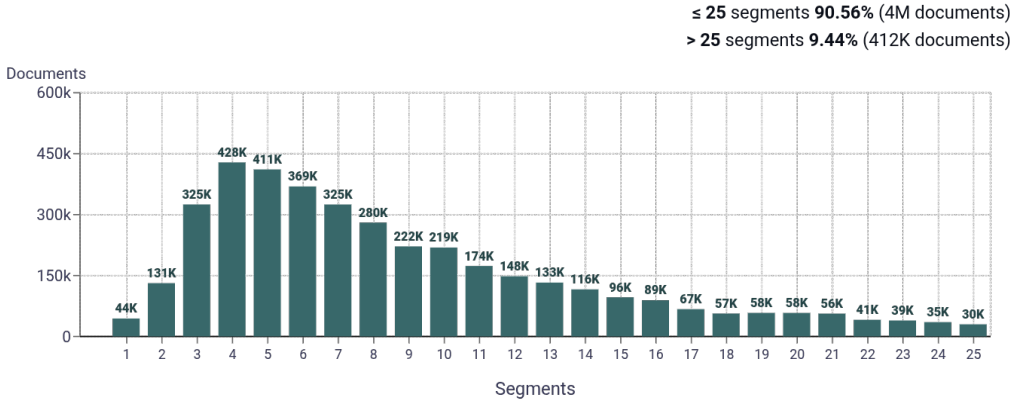
Domain	Docs	% of total
com	3M	67.74%
in	612K	14.03%
net	486K	11.14%
org	101K	2.31%
news	75K	1.72%
gov.in	24K	0.55%
today	23K	0.52%
live	13K	0.31%
co.in	8.7K	0.20%
zone	7.4K	0.17%

Register labels

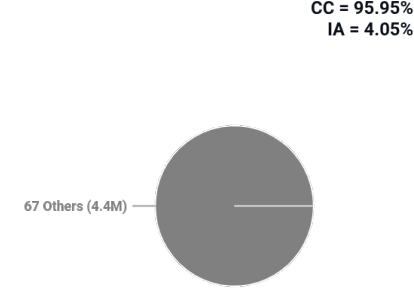


MT:5.9% | 258K Documents

Documents size (in segments) ⓘ

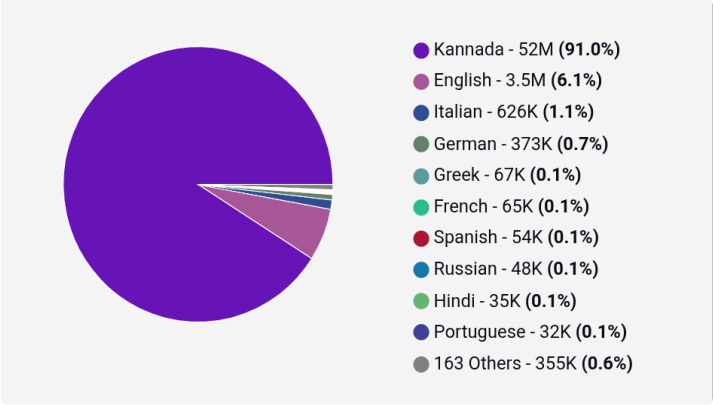


Document collections

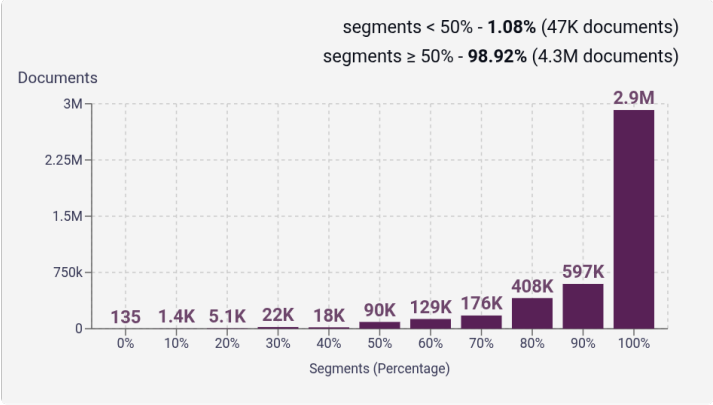


Language Distribution

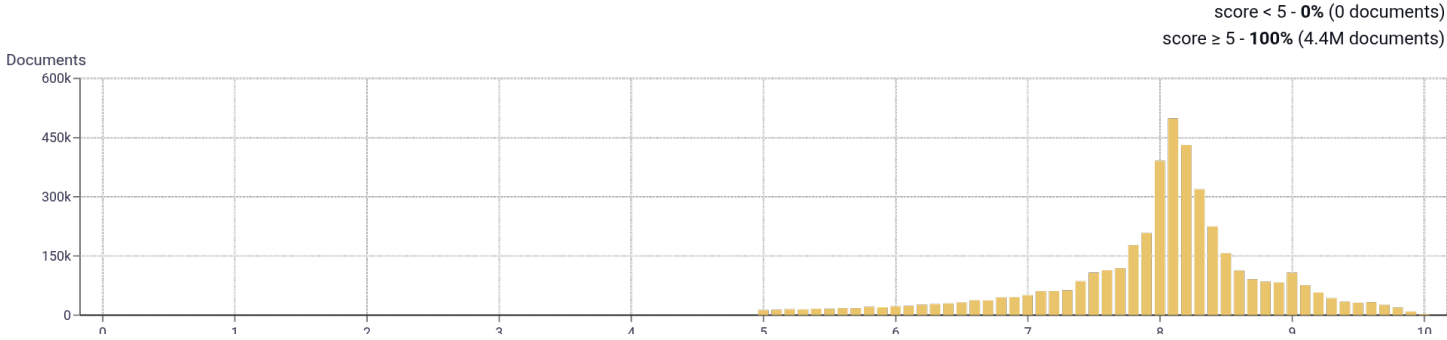
Number of segments in the Kannada corpus



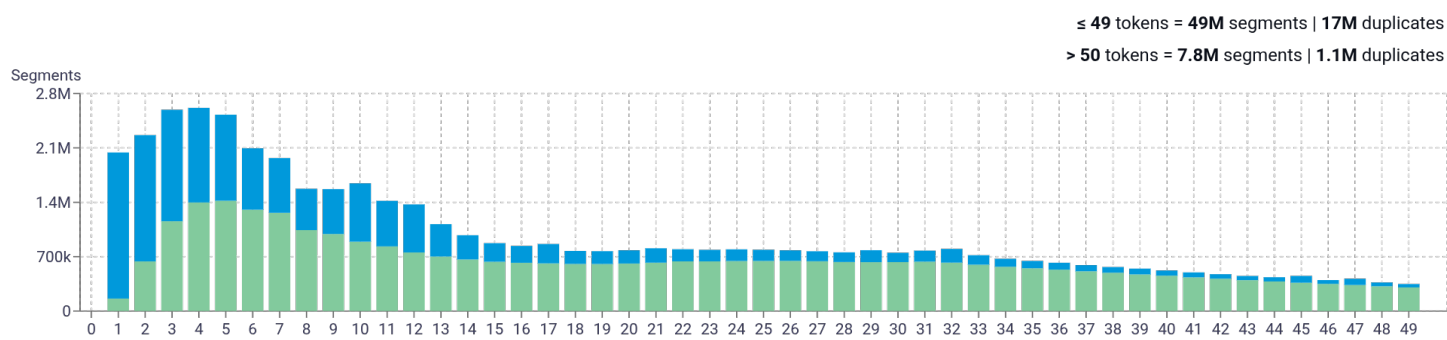
Percentage of segments in Kannada inside documents



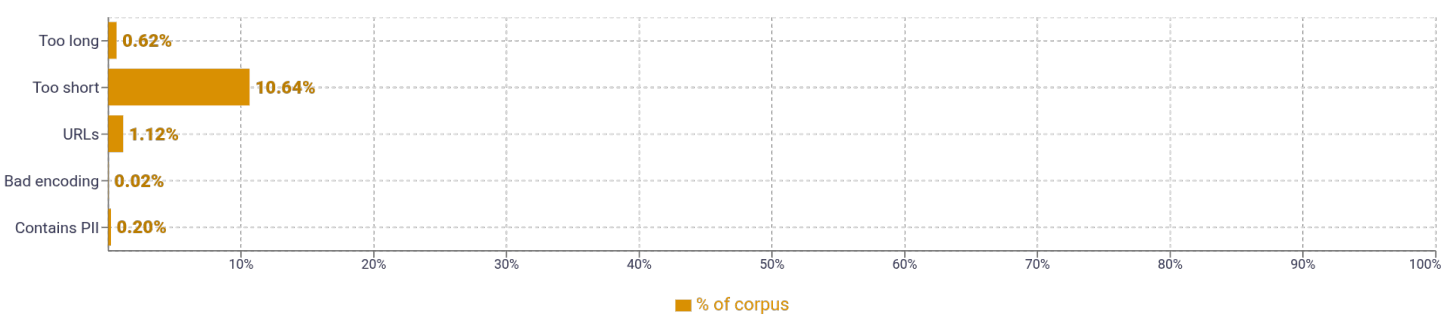
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>ಬೆಂಗಳೂರು 1,238,305</div> <div>ಕಾಂಗ್ರೆಸ್ 1,189,111</div> <div>ಬಿಜೆಪಿ 1,155,560</div> <div>ಇಂದು 1,150,926</div> <div>ವಿರುದ್ಧ 1,145,433</div>	
2	<div>ಇದನ್ನೂ ಓದಿ 335,821</div> <div>ಕೋಟಿ ರೂ 238,808</div> <div>read more 232,484</div> <div>ನರೇಂದ್ರ ಮೋದಿ 217,996</div> <div>ಭೇಟಿ ನೀಡಿ 213,994</div>	
3	<div>ಆಪ್ ಡೌನ್‌ಲೋಡ್ ಮಾಡಿಕೊಳ್ಳಿ 156,211</div> <div>ಪ್ರಜಾವಾಣಿ ಆಪ್ ಡೌನ್‌ಲೋಡ್ 151,999</div> <div>ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಪ್ 151,846</div> <div>ಪ್ರಜಾವಾಣಿ ಫೇಸ್‌ಬುಕ್ ಪುಟವನ್ನು ಫಾಲೋ 151,846</div> <div>ತಾಜಾ ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ 151,846</div>	
4	<div>ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಪ್ ಡೌನ್‌ಲೋಡ್ 151,846</div> <div>ಪ್ರಜಾವಾಣಿ ಆಪ್ ಡೌನ್‌ಲೋಡ್ ಮಾಡಿಕೊಳ್ಳಿ 151,846</div> <div>ತಾಜಾ ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಪ್ 151,846</div> <div>ಚಲಿಸಲಿರುವ ರೈಲುಗಳ ಸಂಪೂರ್ಣ ಪಟ್ಟಿ 86,969</div> <div>ಮಾಹಿತಿ ಪಡೆಯಲು ಪ್ರಜಾವಾಣಿ ಟೆಲಿಗ್ರಾಂ 83,215</div>	
5	<div>ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಪ್ ಡೌನ್‌ಲೋಡ್ 151,846</div> <div>ತಾಜಾ ಸುದ್ದಿಗಳಿಗಾಗಿ ಪ್ರಜಾವಾಣಿ ಆಪ್ ಡೌನ್‌ಲೋಡ್ 151,846</div> <div>ಮಾಹಿತಿ ಪಡೆಯಲು ಪ್ರಜಾವಾಣಿ ಟೆಲಿಗ್ರಾಂ ಚಾನೆಲ್ 83,215</div> <div>ಪಡೆಯಲು ಪ್ರಜಾವಾಣಿ ಟೆಲಿಗ್ರಾಂ ಚಾನೆಲ್ ಸೇರಿಕೊಳ್ಳಿ 83,215</div> <div>ತಾಜಾ ಮಾಹಿತಿ ಪಡೆಯಲು ಪ್ರಜಾವಾಣಿ ಟೆಲಿಗ್ರಾಂ 83,215</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 *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>, , , 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				