HPLT Analytics report

HPLTAnalytics

General overview

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

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%
publictv.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

- HI - 1.0%
- ID - 0.1%
- IN - 5.8%
- IP - 2.9%
- LY - 0.3%
- MIX - 0.8%
- NA - 72.9%
- OP - 4.5%
- SP - 0.2%
- UNK - 11.5%

- HI_other - 0.7%
- HI_re - 0.3%
- ID_other - 0.1%
- IN_dtp - 1.8%
- IN_en - 0.6%
- IN_fi - 0.0%
- IN_lt - 0.0%
- IN_other - 3.4%
- IN_ra - 0.0%
- IP_ds - 1.4%
- IP_other - 1.5%
- LY_other - 0.3%

- MIX - 0.8%
- NA_nb - 1.0%
- NA_ne - 66.4%
- NA_other - 2.7%
- NA_sr - 2.8%
- OP_av - 0.8%
- OP_ob - 0.6%
- OP_other - 1.6%
- OP_rs - 1.0%
- OP_rv - 0.6%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 11.5%
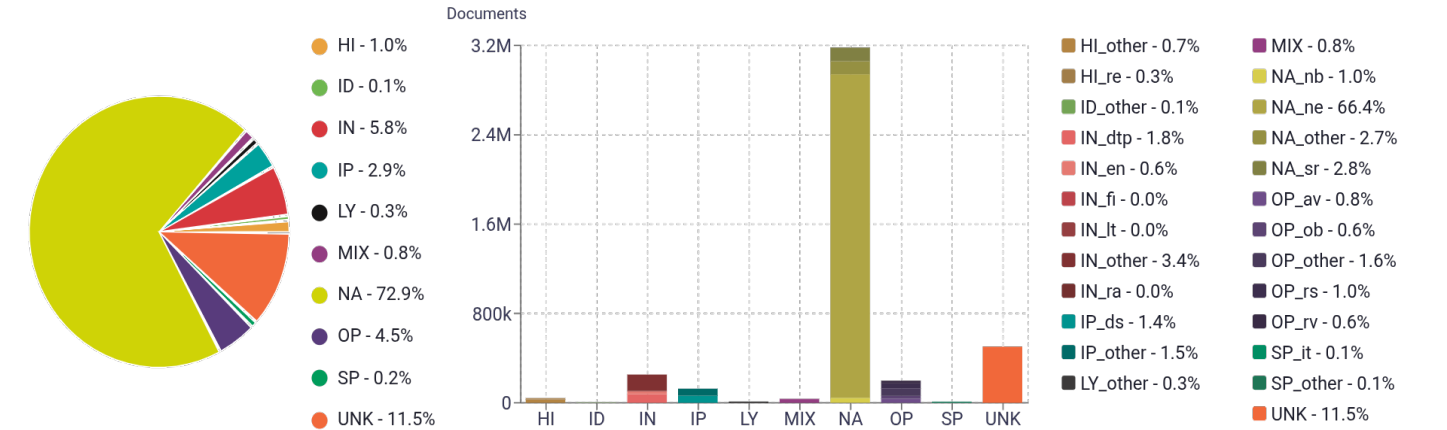
🤖 **MT**:5.9% | 258K Documents

Documents size (in segments) ⓘ

≤ 25 segments **90.56%** (4M documents)
> 25 segments **9.44%** (412K documents)

Document collections

**CC = 95.95%**
IA = 4.05%

67 Others (4.4M)
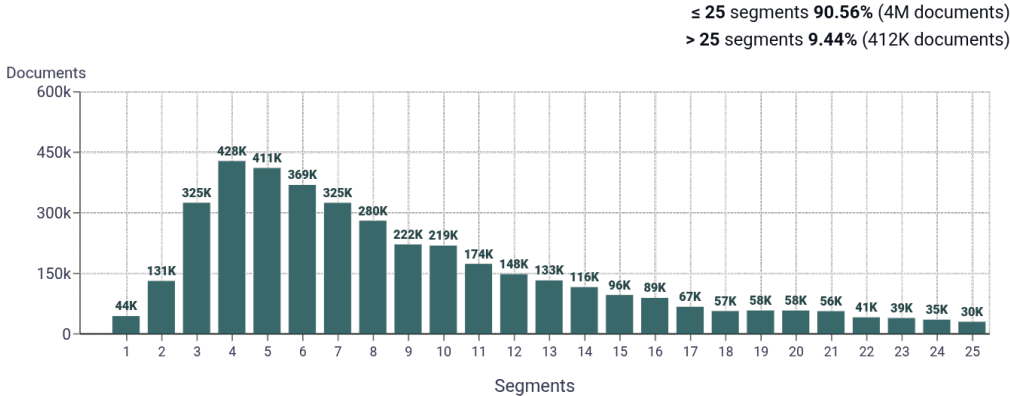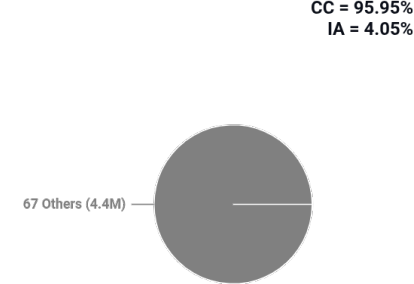
Language Distribution

Number of segments in the Kannada (kn) corpus

- Kannada (kn) - 52M **(91.0%)**
- English (en) - 3.5M **(6.1%)**
- Italian (it) - 626K **(1.1%)**
- German (de) - 373K **(0.7%)**
- Greek (el) - 67K **(0.1%)**
- French (fr) - 65K **(0.1%)**
- Spanish (es) - 54K **(0.1%)**
- Russian (ru) - 48K **(0.1%)**
- Hindi (hi) - 35K **(0.1%)**
- Portuguese (pt) - 32K **(0.1%)**
- 163 Others - 355K **(0.6%)**
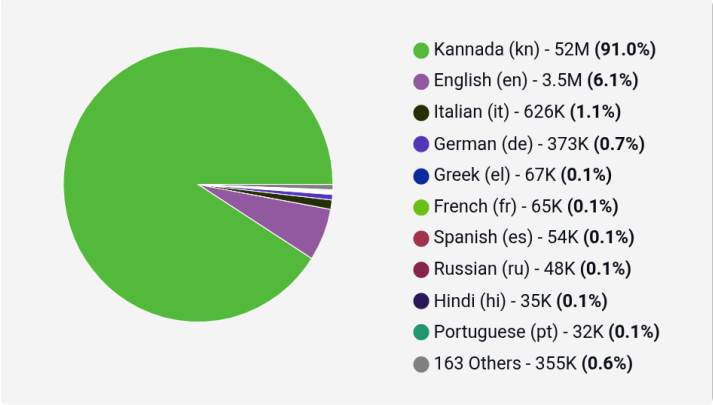
Percentage of segments in Kannada (kn) inside documents
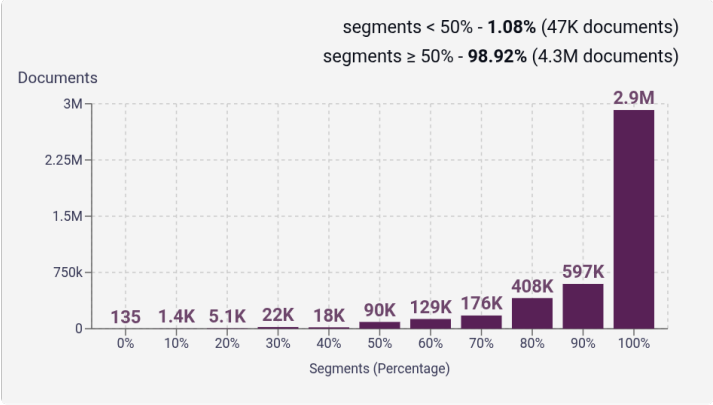
segments < 50% - **1.08%** (47K documents)
segments ≥ 50% - **98.92%** (4.3M documents)

Documents

135	1.4K	5.1K	22K	18K	90K	129K	176K	408K	597K	2.9M
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4.4M documents)

Documents

Segment length distribution by token

≤ **49** tokens = **49M** segments | **17M** duplicates
> **50** tokens = **7.8M** segments | **1.1M** duplicates

Segments

Segment noise distribution

- Too long — **0.62%**
- Too short — **10.64%**
- URLs — **1.12%**
- Bad encoding — **0.02%**
- Contains PII — **0.20%**

■ % of corpus
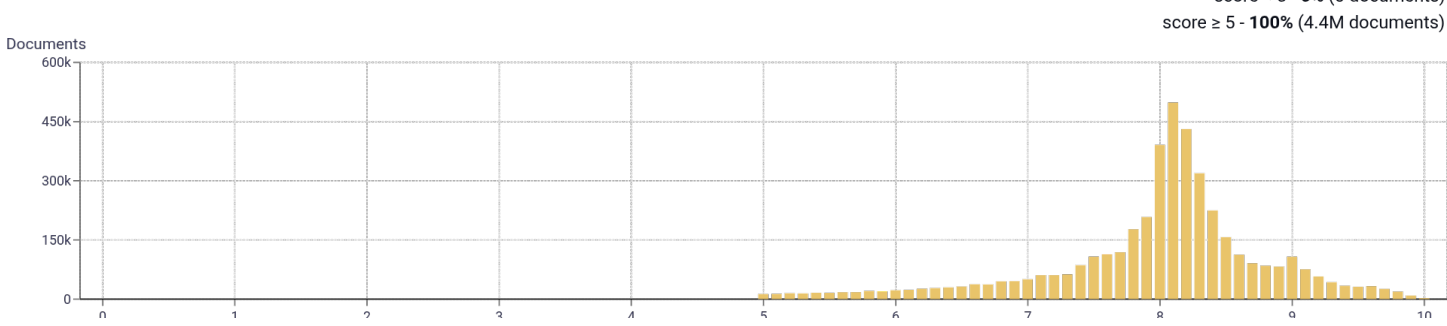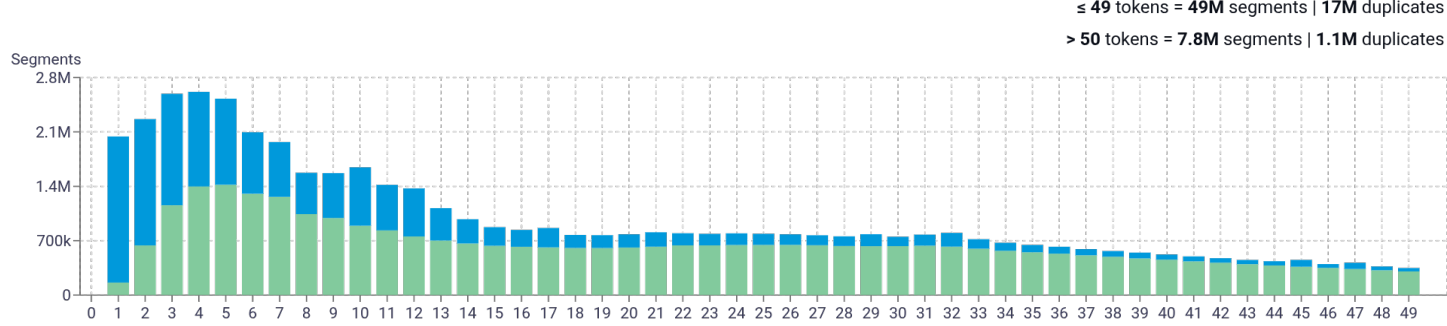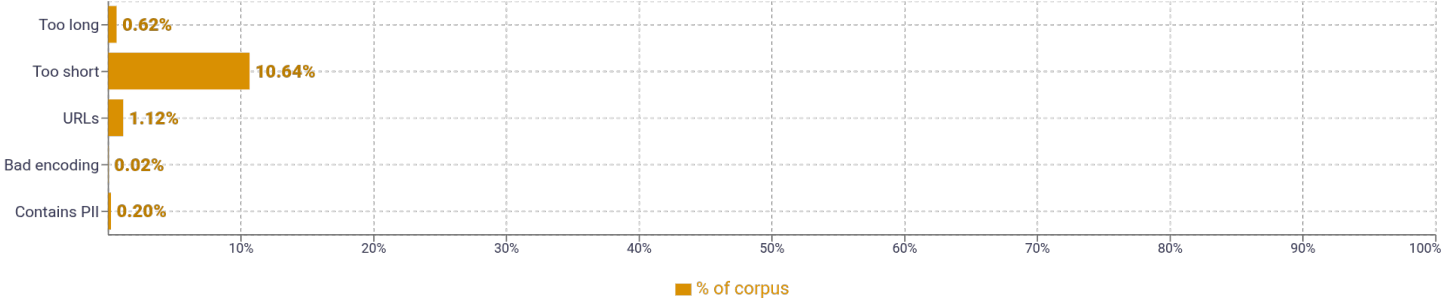
Frequent n-grams

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

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 or 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				