HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-hne_Deva	9/17/2025	Chhattisgarhi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,322	95,601	80,942 (84.67 %)	4.8M	20,425,315	48.12 MB

Top 10 domains

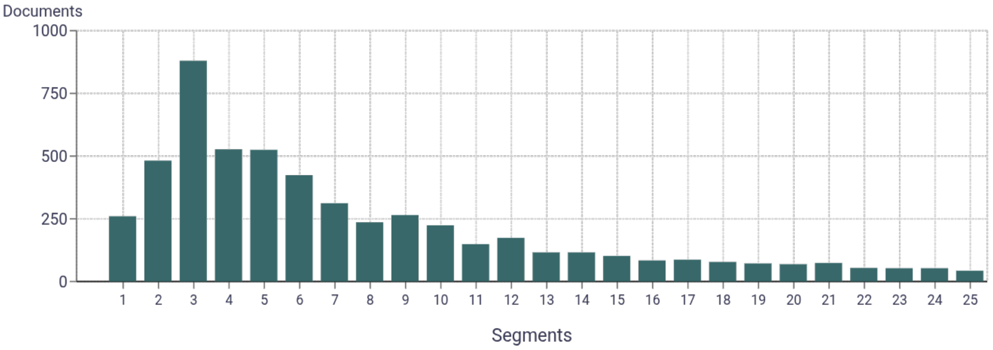
Domain	Docs	% of total
gurturgoth.com	1.4K	22.68%
jayjohar.com	1.3K	20.91%
anjor.online	567	8.97%
ruralindiaonlin...	565	8.94%
blogspot.com	406	6.42%
news36live.com	289	4.57%
news18.com	278	4.40%
biblica.com	181	2.86%
patrika.com	172	2.72%
surta.in	68	1.08%

Top 10 TLDs

Domain	Docs	% of total
com	4.7K	74.14%
org	621	9.82%
online	567	8.97%
in	214	3.39%
net	104	1.65%
co.in	66	1.04%
com.ua	9	0.14%
info	6	0.09%
club	6	0.09%
page	4	0.06%

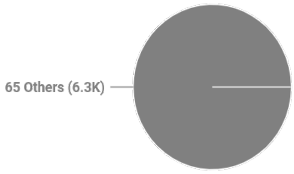
Documents size (in segments) ⓘ

≤ **25** segments **86.35%** (5.5K documents)
> 25 segments **13.65%** (863 documents)



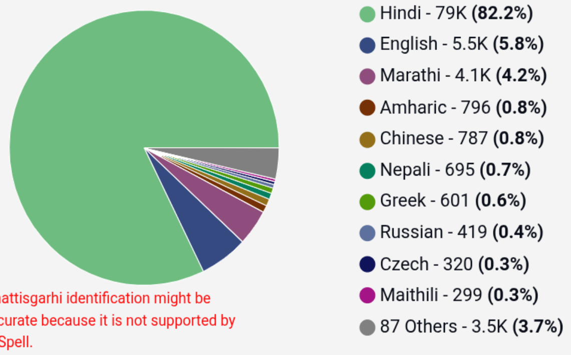
Document collections

CC = 95.71%
IA = 4.29%



Language Distribution

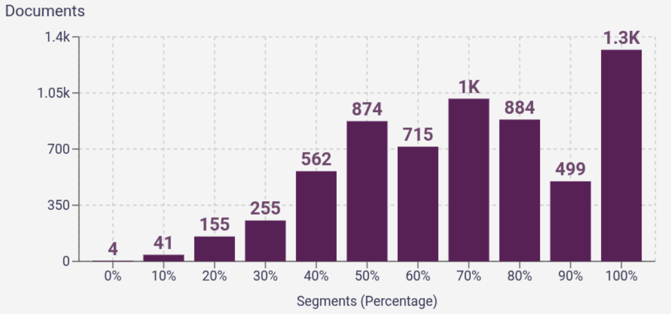
Number of segments in the Chhattisgarhi corpus



*Chhattisgarhi identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Chhattisgarhi inside documents

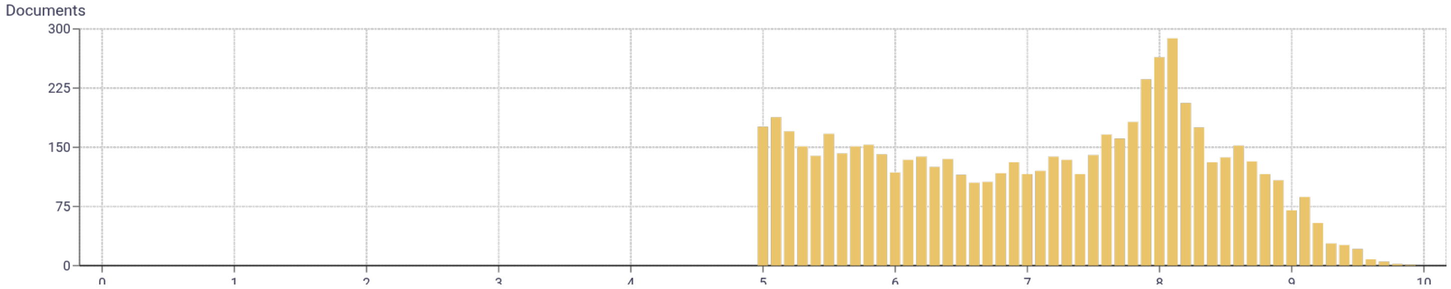
segments < 50% - **16.09%** (1K documents)
segments ≥ 50% - **83.91%** (5.3K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

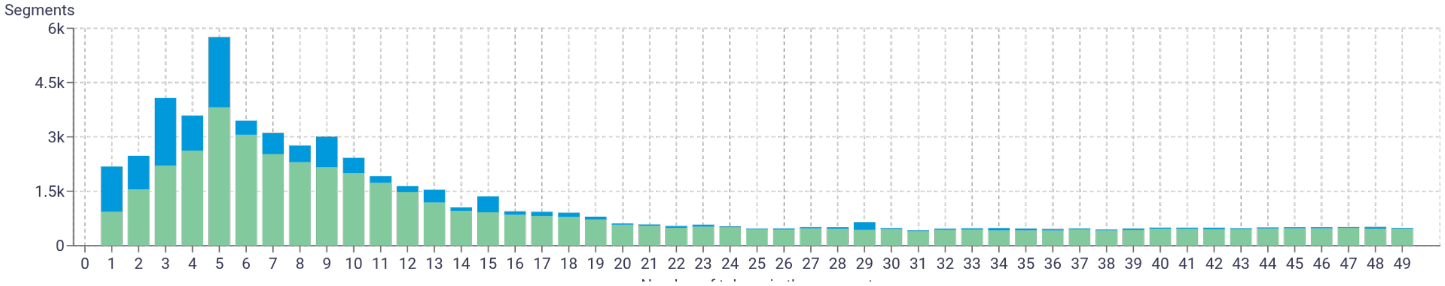
score ≥ 5 - **100%** (6.3K documents)



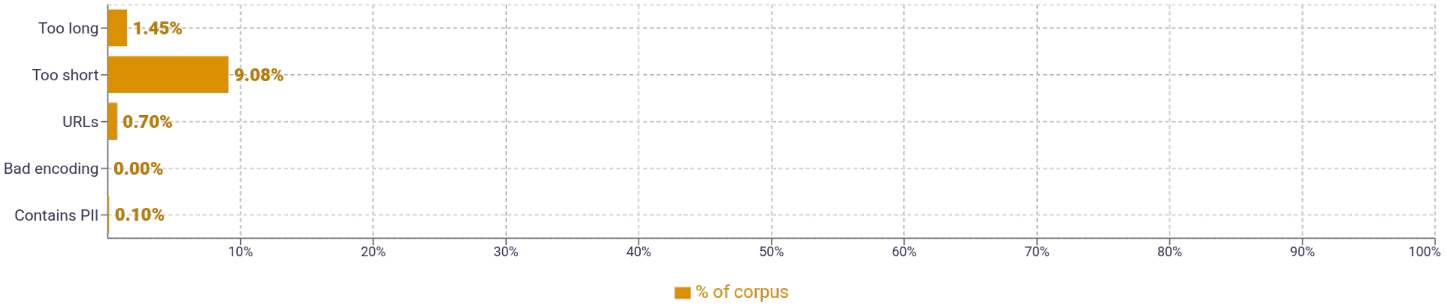
Segment length distribution by token

≤ **49** tokens = **59K** segments | **13K** duplicates

> 50 tokens = **36K** segments | **2.1K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	मन | 71,039 ले | 69,560 ह | 58,514 हे | 51,183 ला | 47,417
2	वो ह | 10,590 मन ला | 6,456 वो मन | 4,837 लोगन मन | 3,953 मन ले | 2,755
3	keyword keyword keyword | 2,130 वो ह कहिथे | 1,520 वो मन ला | 1,095 item item item | 618 निर्मल कुमार साहू | 558
4	keyword keyword keyword keyword | 2,118 item item item item | 614 be the first one | 262 item selected item selected | 245 selected item selected item | 244
5	keyword keyword keyword keyword keyword | 2,108 item item item item item | 610 item selected item selected item | 244 selected item selected item selected | 240 हमर भाखा ल आसानी ले | 133

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				