

General overview

Corpus	Date	Language
hplt-v3-ory_Orya	9/24/2025	Odia

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,297,982	9,436,011	7,569,205 (80.22 %)	338M	2,203,269,520	5.43 GB

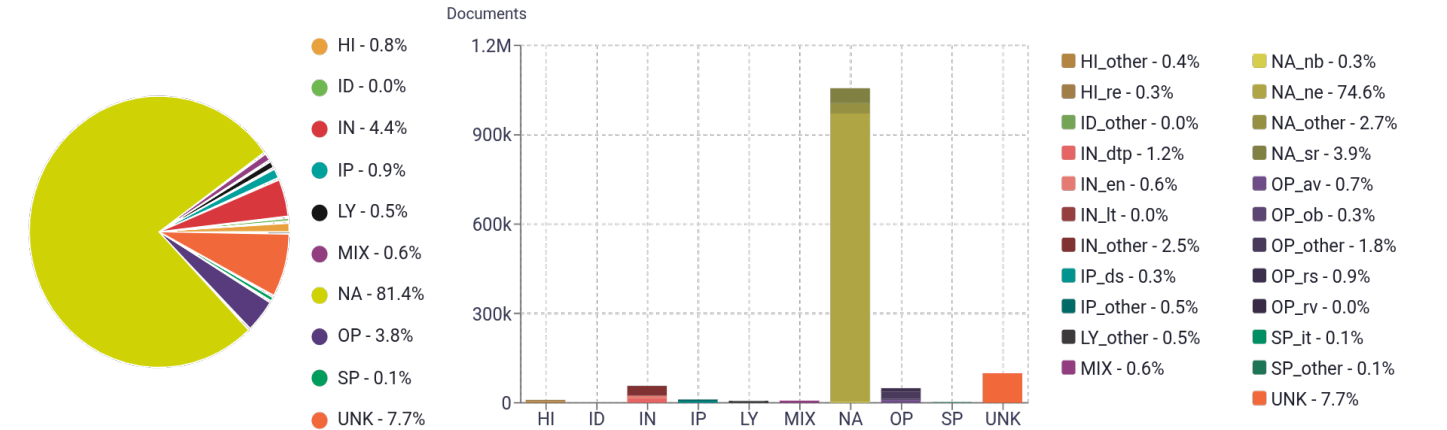
Top 10 domains

Domain	Docs	% of total
sambad.in	134K	10.32%
dharitri.com	82K	6.35%
odishareporter.in	77K	5.95%
kanaknews.com	66K	5.07%
samajalive.in	63K	4.88%
prameyanews7.com	55K	4.25%
knewsodisha.com	30K	2.31%
prameya.com	25K	1.96%
news18.com	24K	1.84%
odiapua.com	23K	1.77%

Top 10 TLDs

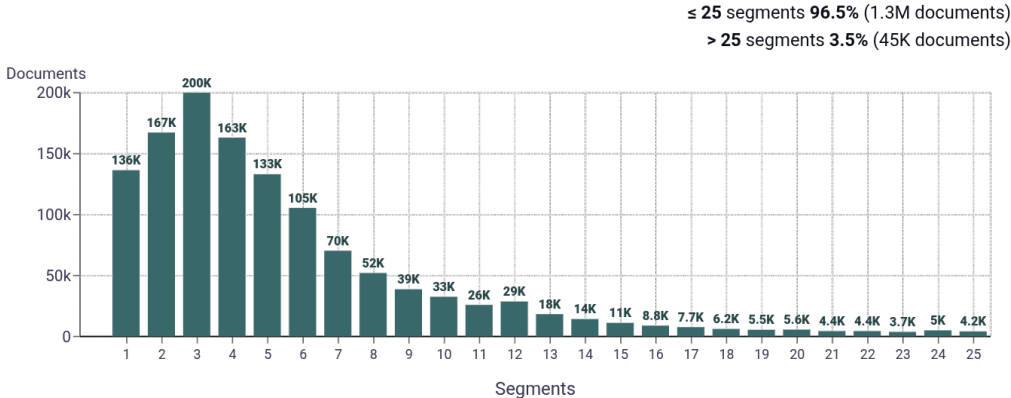
Domain	Docs	% of total
com	762K	58.68%
in	456K	35.12%
org	26K	1.99%
me	18K	1.38%
live	10K	0.79%
info	5.6K	0.43%
net	4.8K	0.37%
co	4K	0.31%
co.in	2.4K	0.18%
gov.in	2.4K	0.18%

Register labels

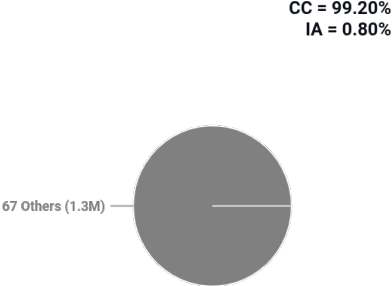


MT:3.1% | 40K Documents

Documents size (in segments)

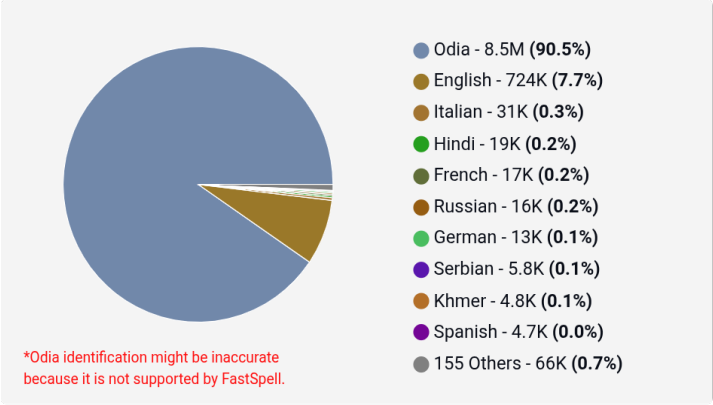


Document collections

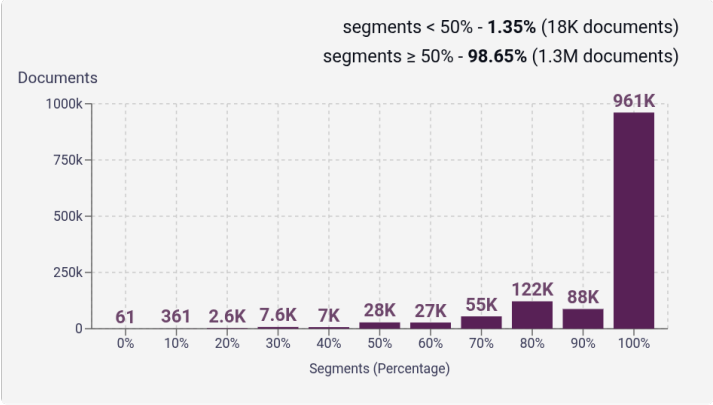


Language Distribution

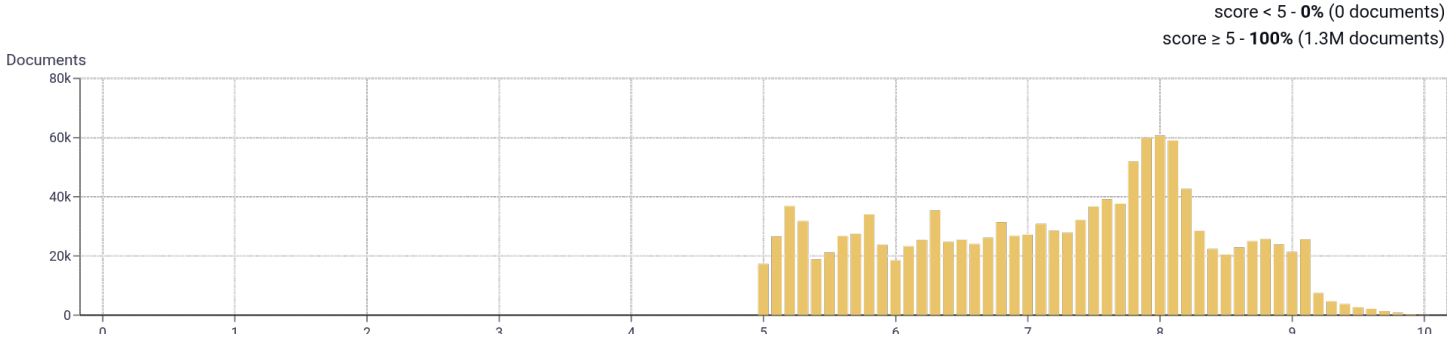
Number of segments in the Odia corpus



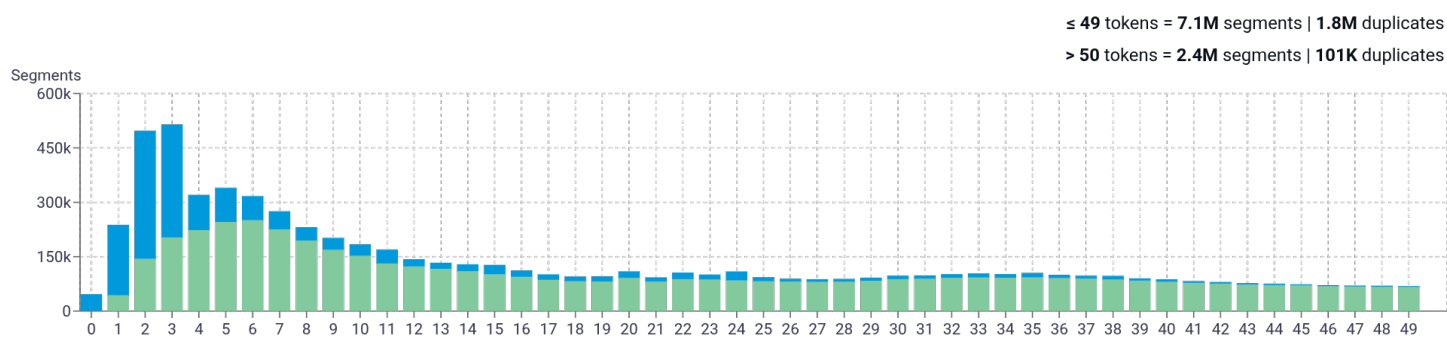
Percentage of segments in Odia inside documents



Distribution of documents by document score

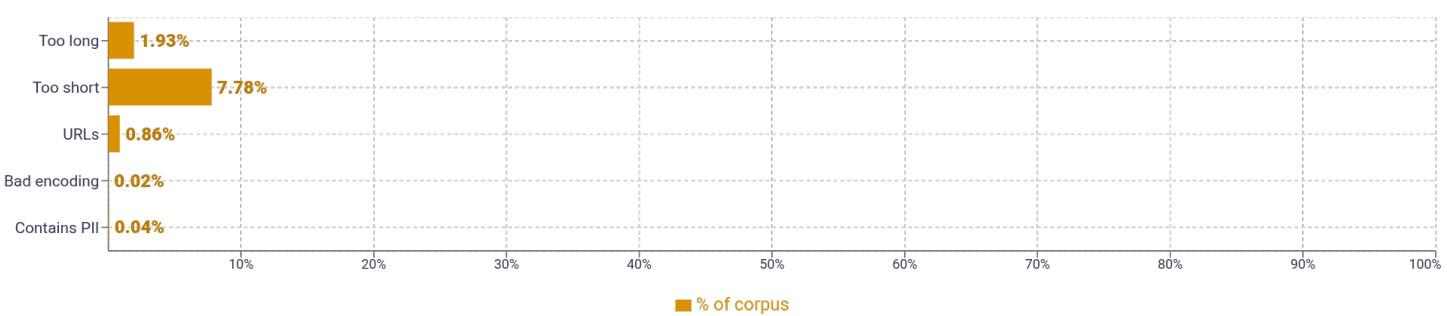


Segment length distribution by token



≤ 49 tokens = 7.1M segments | 1.8M duplicates
> 50 tokens = 2.4M segments | 101K duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ଏକ 1,722,172 ସେ 1,604,057 କରିବା 1,176,132 ବୋଲି 1,065,588 ଯେ 890,091	
2	କହିଛନ୍ତି ଯେ 182,145 read more 98,114 comments are 92,306 are closed 92,159 ଗଲ୍ଲୁ ସରକାର 91,838	
3	comments are closed 92,152 ଥିବାକ ଖବର ଦେଖନ୍ତୁ 52,638 pic twitter com 40,584 ଯୁଖିଯୁଖିଲା ନବୀନ ପଟ୍ଟନାୟକ 36,685 ସେ କହିଛନ୍ତି ଯେ 29,474	
4	ପ୍ରେସାଇଡେନ୍ଟ ସବୁଠାରୁ ନିର୍ଭରଯୋଗ୍ୟ ଓଡ଼ିଆ 13,916 ହୁଅନ୍ତୁ ଆଜିର ସର୍ବଶେଷ ଖବର 13,916 ସର୍ବଶେଷ ଖବର ଲାଇଭ୍ ରୁୟାଲ୍ 13,916 ସବୁଠାରୁ ନିର୍ଭରଯୋଗ୍ୟ ଓଡ଼ିଆ ଖବର 13,916 ଲାଇଭ୍ ରୁୟାଲ୍ ଅପଡେଟ୍ ରୁୟାଲ୍ 13,916	
5	ପ୍ରେସାଇଡେନ୍ଟ ସବୁଠାରୁ ନିର୍ଭରଯୋଗ୍ୟ ଓଡ଼ିଆ ଖବର 13,916 ହୁଅନ୍ତୁ ଆଜିର ସର୍ବଶେଷ ଖବର ଲାଇଭ୍ 13,916 ସର୍ବଶେଷ ଖବର ଲାଇଭ୍ ରୁୟାଲ୍ ଅପଡେଟ୍ 13,916 ସବୁଠାରୁ ନିର୍ଭରଯୋଗ୍ୟ ଓଡ଼ିଆ ଖବର ପଢନ୍ତୁ 13,916 ବ୍ରେକିଙ୍ଗ୍ ରୁୟାଲ୍ ପଢିବାରେ ପରଥମ ହୁଅନ୍ତୁ 13,916	

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				