

General overview

Corpus	Date	Language
hplt-v3-mag_Deva	9/17/2025	Magahi (mag)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
513	75,605	53,958 (71.37 %)	1.1M	4,670,579	11.24 MB

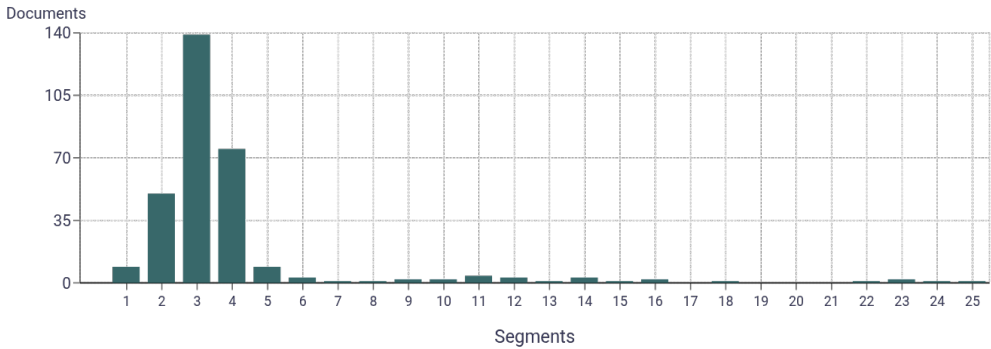
Top 10 domains

Domain	Docs	% of total
khabarlahariya.org	256	49.90%
blogspot.com	186	36.26%
blogspot.in	22	4.29%
wikipedia.org	12	2.34%
gotquestions.org	9	1.75%
wikimedia.org	7	1.36%
kavitakosh.org	2	0.39%
unicode.org	1	0.19%
teachersofbihar...	1	0.19%
sinstongasket.com	1	0.19%

Top 10 TLDs

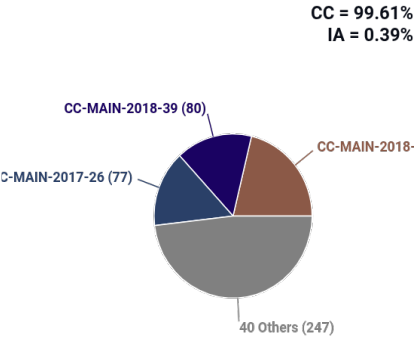
Domain	Docs	% of total
org	290	56.53%
com	198	38.60%
in	23	4.48%
info	1	0.19%
buzz	1	0.19%

Documents size (in segments) ⓘ



≤ 25 segments **60.62%** (311 documents)  
> 25 segments **39.38%** (202 documents)

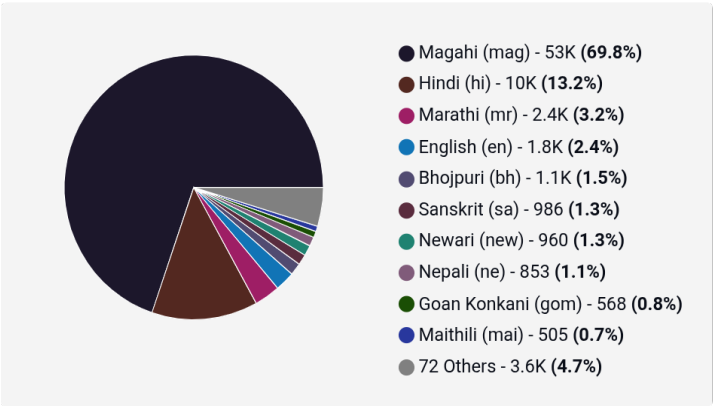
Document collections



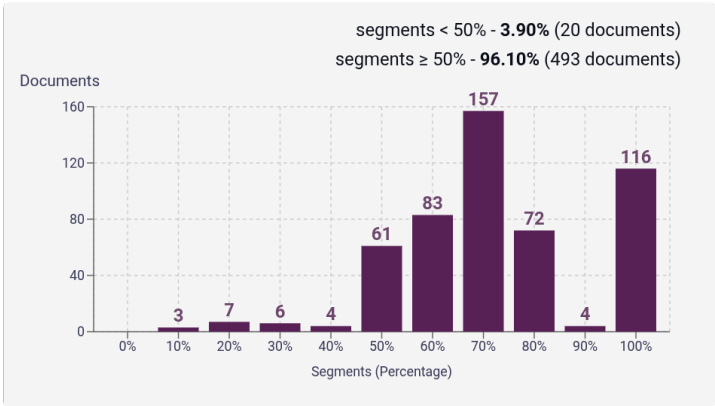
CC = **99.61%**  
IA = **0.39%**

Language Distribution

Number of segments in the Magahi (mag) corpus

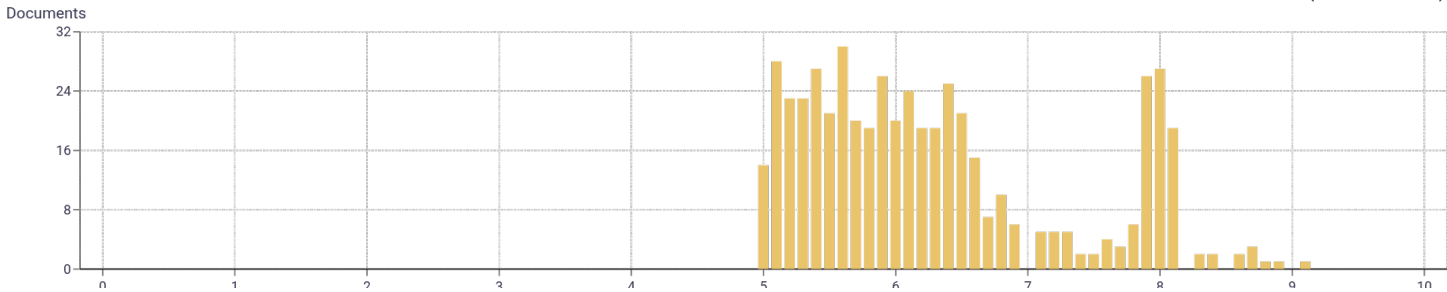


Percentage of segments in Magahi (mag) inside documents



segments < 50% - **3.90%** (20 documents)  
segments ≥ 50% - **96.10%** (493 documents)

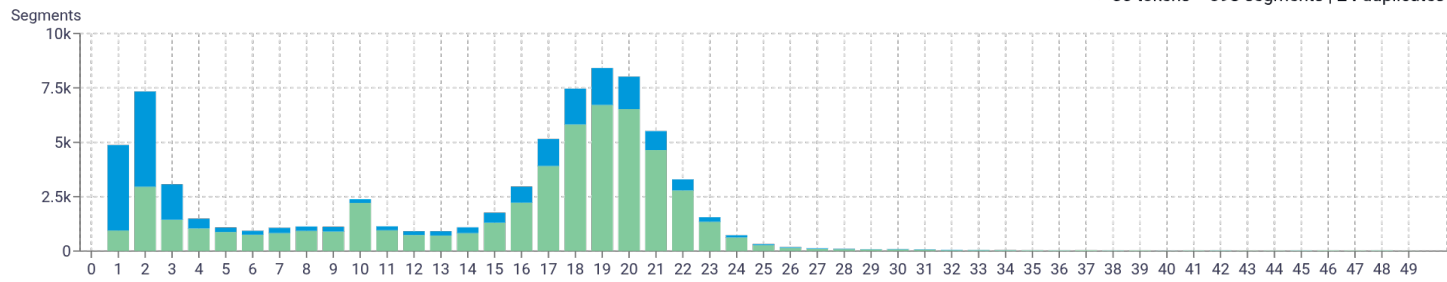
Distribution of documents by document score



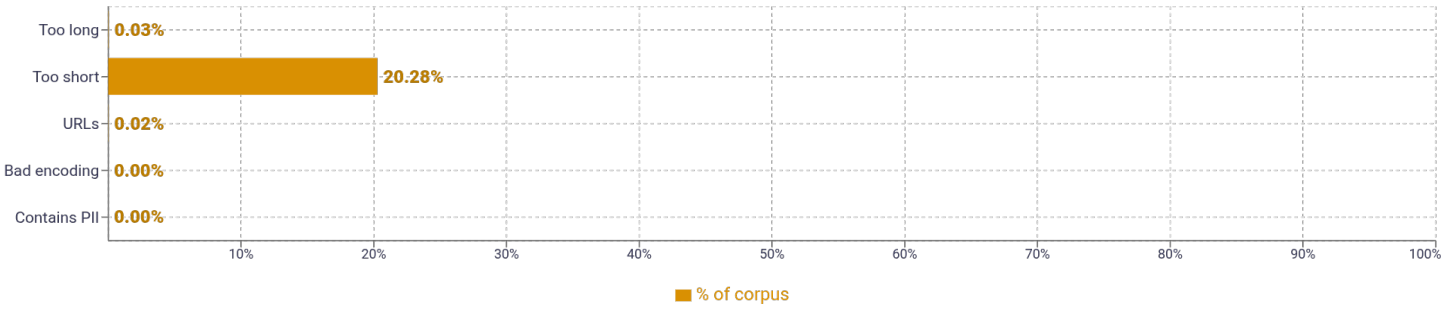
score < 5 - **0%** (0 documents)  
score ≥ 5 - **100%** (513 documents)

Segment length distribution by token

≤ 49 tokens = 75K segments | 22K duplicates  
> 50 tokens = 693 segments | 24 duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	आउ   25,369    में   22,554    हइ   15,882    हलइ   15,024    ऊ   12,038	📄
2	ट्रांसजेंडर व्यक्ति   2,392    एक ट्रांसजेंडर   2,391    रहले हल   1,928    जा हइ   1,384    करऽ हलइ   1,370	📄
3	एक ट्रांसजेंडर व्यक्ति   2,389    व्यक्ति के माता   2,152    जा सकऽ हइ   416    जा सकऽ हलइ   191    व्यक्ति के रिश्तेदार   182	📄
4	ट्रांसजेंडर व्यक्ति के माता   2,152    ट्रांसजेंडर व्यक्ति के रिश्तेदार   182    कइल जा सकऽ हइ   115    चूंकि चूंकि चूंकि चूंकि   92    कहल जा सकऽ हइ   54	📄
5	एक ट्रांसजेंडर व्यक्ति के माता   2,151    एक ट्रांसजेंडर व्यक्ति के रिश्तेदार   180    चूंकि चूंकि चूंकि चूंकि चूंकि   90    एक ट्रांसजेंडर व्यक्ति का सहोदर   27    नयँ कइल जा सकऽ हइ   24	📄

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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>, <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	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				