# HPLT Analytics report

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-mag_Deva | 9/17/2025 | Magahi |

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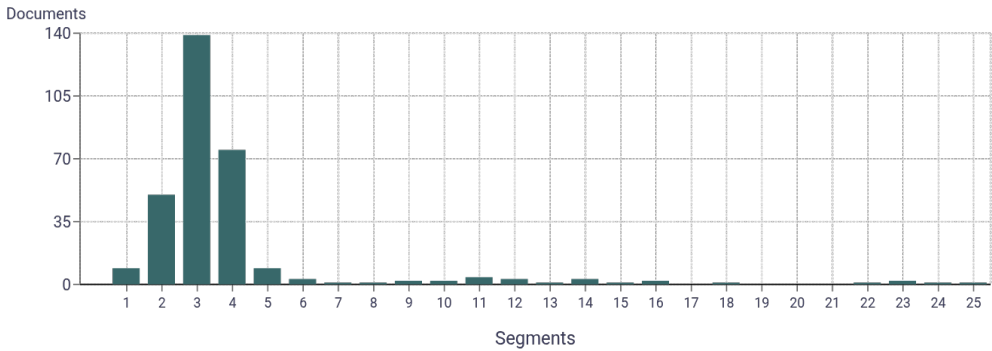
≤ **25** segments **60.62%** (311 documents)
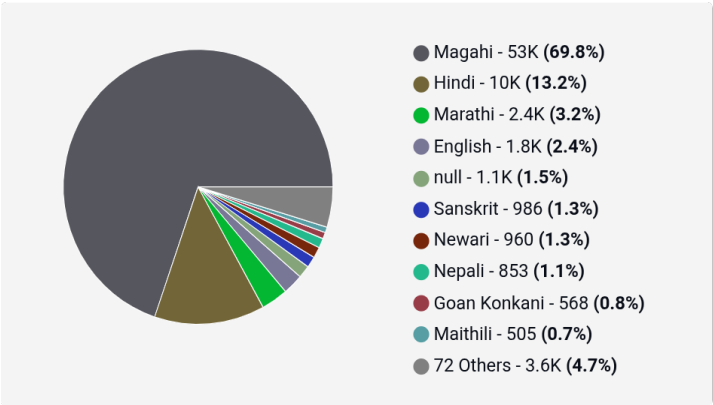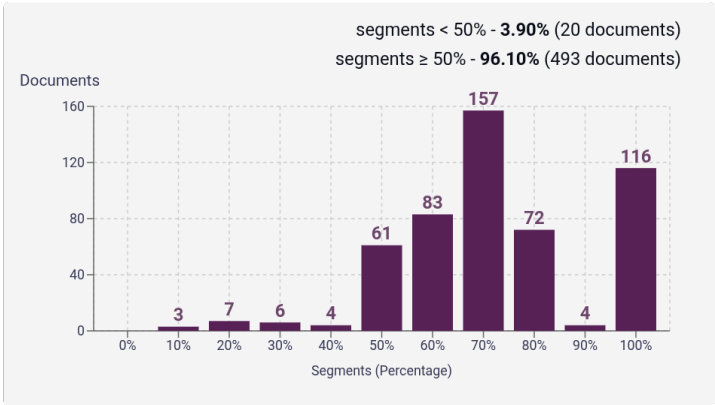> **25** segments **39.38%** (202 documents)



## Document collections

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



CC-MAIN-2018-39 (80)
CC-MAIN-2018-
C-MAIN-2017-26 (77)
40 Others (247)
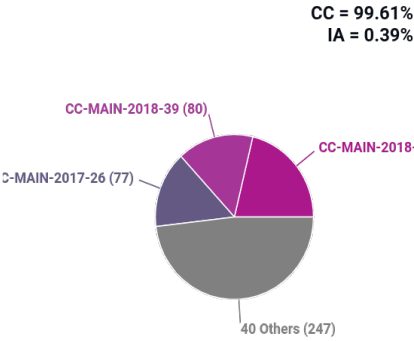
## Language Distribution

### Number of segments in the Magahi corpus



- Magahi - 53K **(69.8%)**
- Hindi - 10K **(13.2%)**
- Marathi - 2.4K **(3.2%)**
- English - 1.8K **(2.4%)**
- null - 1.1K **(1.5%)**
- Sanskrit - 986 **(1.3%)**
- Newari - 960 **(1.3%)**
- Nepali - 853 **(1.1%)**
- Goan Konkani - 568 **(0.8%)**
- Maithili - 505 **(0.7%)**
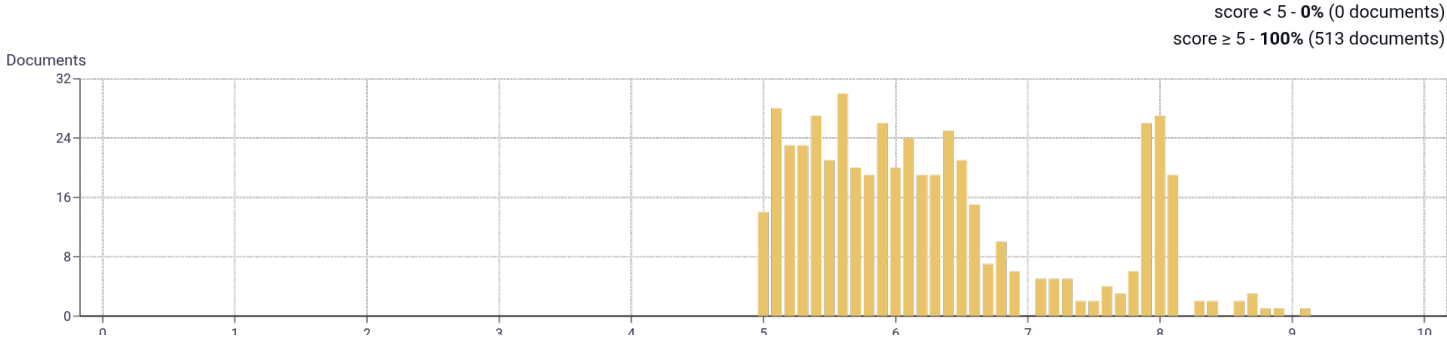- 72 Others - 3.6K **(4.7%)**
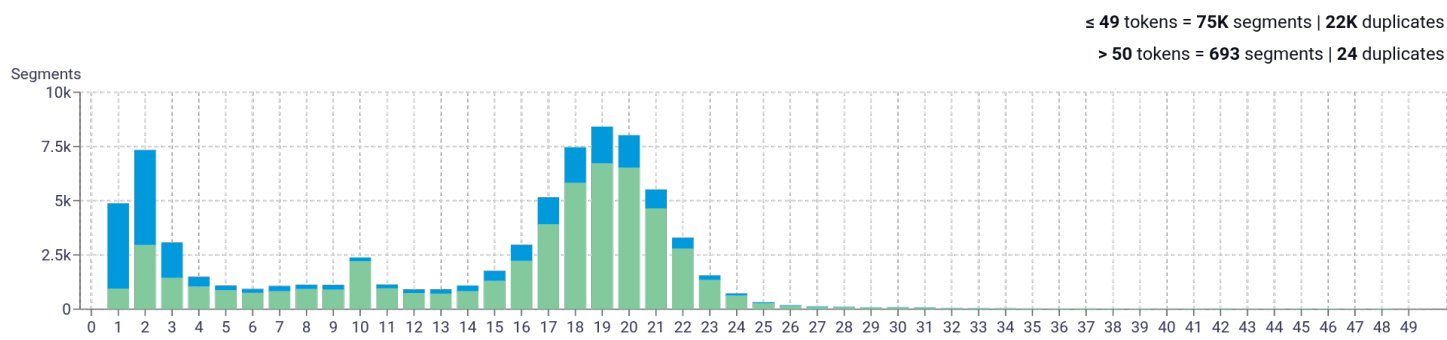
### Percentage of segments in Magahi inside documents

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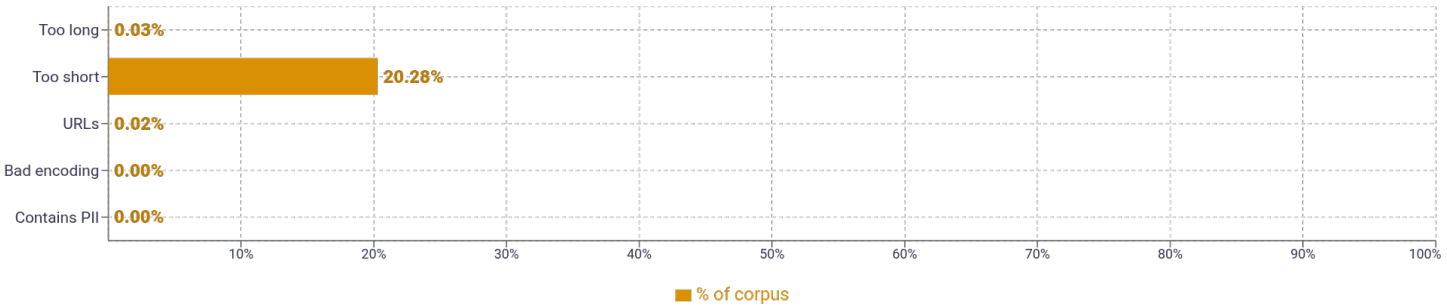segments ≥ 50% - **96.10%** (493 documents)

## Distribution of documents by document score

Documents

32

24

16

8

0

0   1   2   3   4   5   6   7   8   9   10

## Segment length distribution by token

**≤ 49** tokens = **75K** segments | **22K** duplicates

**> 50** tokens = **693** segments | **24** duplicates

Segments
10k

7.5k

5k

2.5k

0

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49

Number of tokens in the segment

## Segment noise distribution

| | |
|---|---|
| Too long | **0.03%** |
| Too short | **20.28%** |
| URLs | **0.02%** |
| Bad encoding | **0.00%** |
| Contains PII | **0.00%** |

10%  20%  30%  40%  50%  60%  70%  80%  90%  100%

■ % of corpus

## 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 *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 | | | | |