# HPLT Analytics report

**HPLT** Analytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-hin_Deva | 9/18/2025 | Hindi |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 36,325,196 | 563,360,834 | 398,763,470 (70.78 %) | 22B | 99,222,596,922 | 230.32 GB |

## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| indiatimes.com | 505K | 1.39% |
| blogspot.com | 493K | 1.36% |
| jagran.com | 469K | 1.29% |
| bhaskar.com | 418K | 1.15% |
| patrika.com | 357K | 0.98% |
| news18.com | 333K | 0.92% |
| amarujala.com | 288K | 0.79% |
| punjabkesari.in | 258K | 0.71% |
| livehindustan.com | 253K | 0.70% |
| india.com | 230K | 0.63% |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| com | 26M | 71.52% |
| in | 6.2M | 16.93% |
| org | 905K | 2.49% |
| co.in | 526K | 1.45% |
| net | 408K | 1.12% |
| page | 377K | 1.04% |
| news | 245K | 0.67% |
| co | 162K | 0.44% |
| ru | 123K | 0.34% |
| xyz | 121K | 0.33% |

## Register labels



- HI - 2.0%
- ID - 0.2%
- IN - 11.3%
- IP - 3.4%
- LY - 0.7%
- MIX - 1.6%
- NA - 63.1%
- OP - 5.3%
- SP - 0.2%
- UNK - 12.2%

- HI_other - 1.6%
- HI_re - 0.4%
- ID_other - 0.2%
- IN_dtp - 3.9%
- IN_en - 0.2%
- IN_fi - 0.0%
- IN_lt - 0.1%
- IN_other - 7.0%
- IN_ra - 0.0%
- IP_ds - 2.4%
- IP_other - 1.0%
- LY_other - 0.7%
- MIX - 1.6%
- NA_nb - 0.6%
- NA_ne - 57.0%
- NA_other - 2.6%
- NA_sr - 2.9%
- OP_av - 1.0%
- OP_ob - 1.5%
- OP_other - 1.4%
- OP_rs - 0.8%
- OP_rv - 0.6%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 12.2%
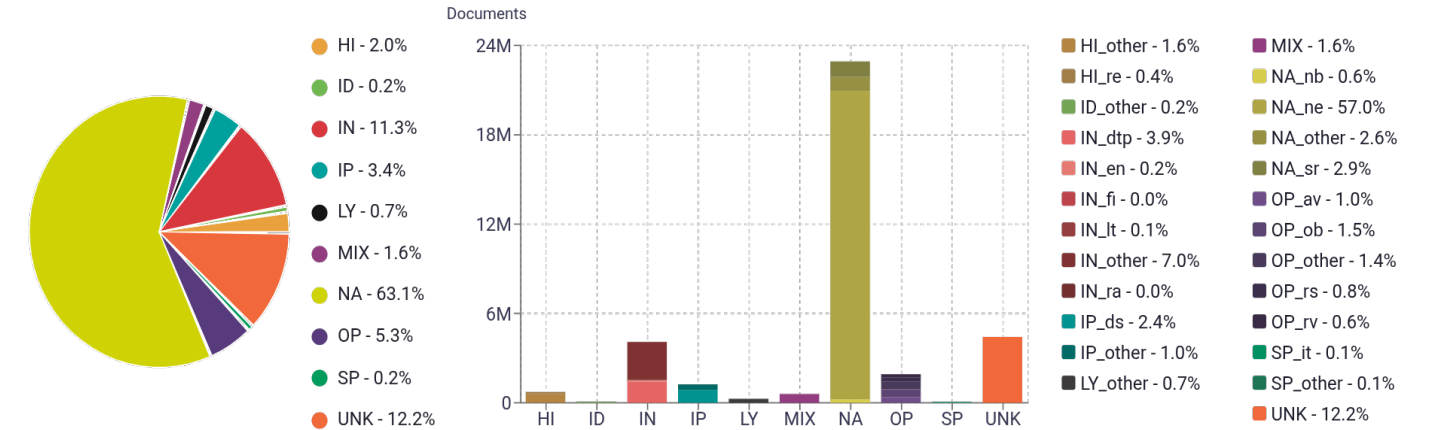
🤖 **MT**:7.0% | 2.5M Documents

## Documents size (in segments) ⓘ

≤ 25 segments **86.39%** (31M documents)
> 25 segments **13.61%** (4.9M documents)



## Document collections

**CC = 95.81%**
**IA = 4.19%**



67 Others (36M)
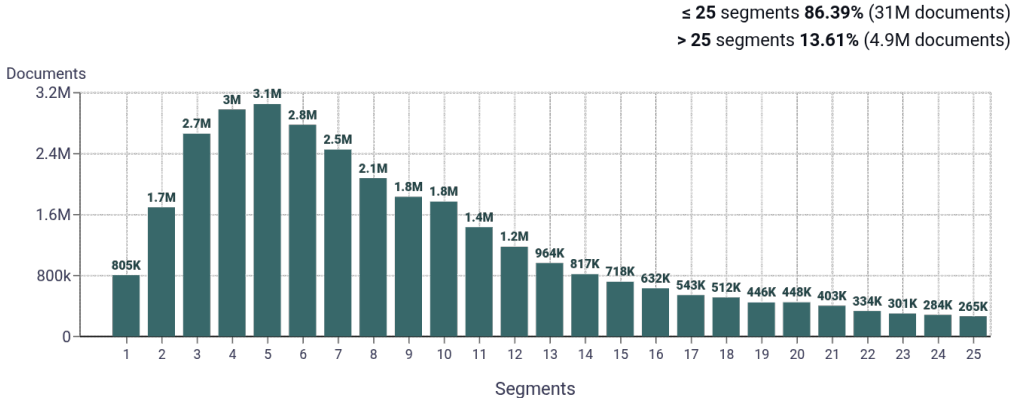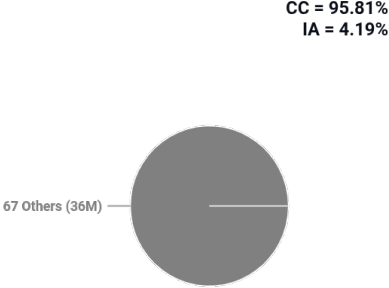
## Language Distribution

### Number of segments in the Hindi corpus



- Hindi - 491M **(87.2%)**
- English - 41M **(7.4%)**
- Nepali - 10M **(1.8%)**
- Italian - 5.2M **(0.9%)**
- Marathi - 2.1M **(0.4%)**
- French - 1.5M **(0.3%)**
- German - 1.4M **(0.3%)**
- Sanskrit - 1.2M **(0.2%)**
- Newari - 1.1M **(0.2%)**
- Spanish - 662K **(0.1%)**
- 165 Others - 7.3M **(1.3%)**
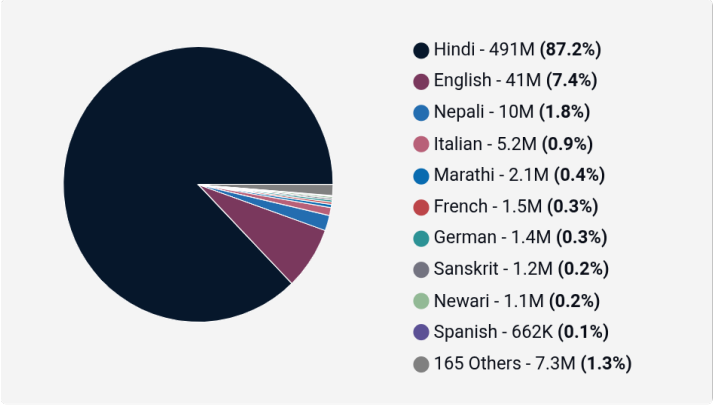
### Percentage of segments in Hindi inside documents
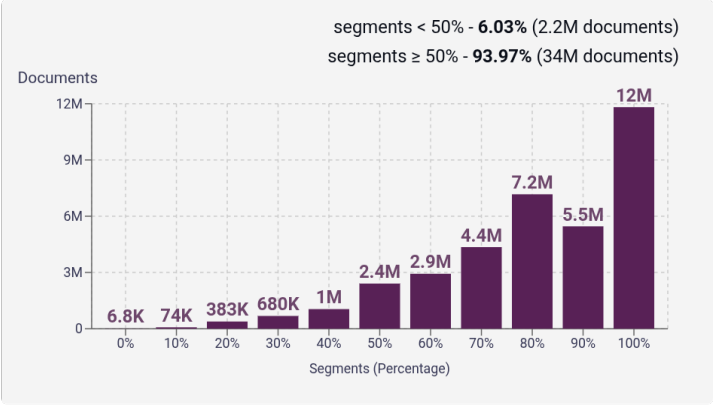
segments < 50% - **6.03%** (2.2M documents)
segments ≥ 50% - **93.97%** (34M documents)



### Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (36M documents)



### Segment length distribution by token

**≤ 49** tokens = **408M** segments | **147M** duplicates
**> 50** tokens = **156M** segments | **18M** duplicates



### Segment noise distribution

- Too long — **0.85%**
- Too short — **8.60%**
- URLs — **1.09%**
- Bad encoding — **0.00%**
- Contains PII — **0.09%**

■ % of corpus
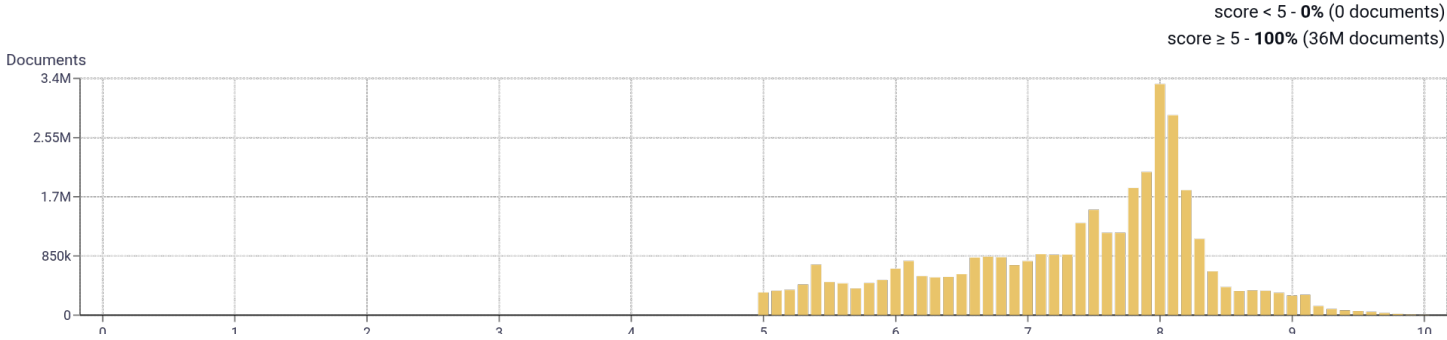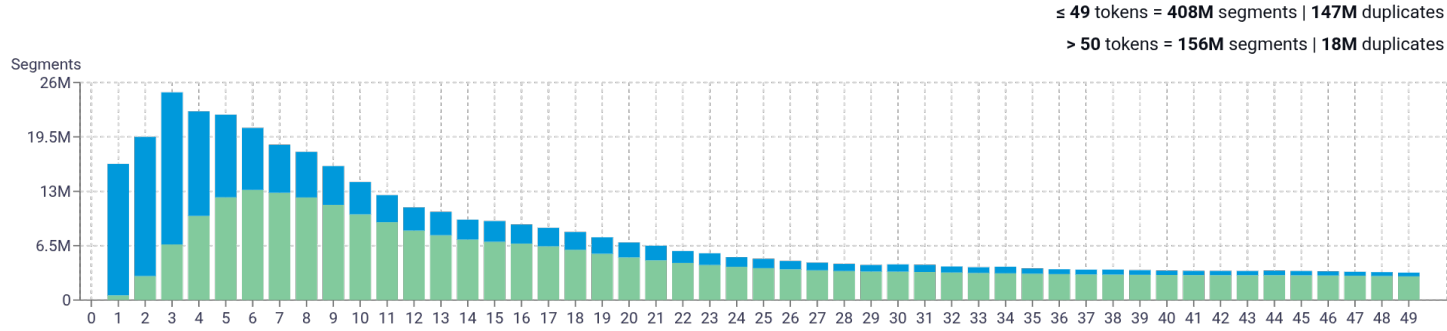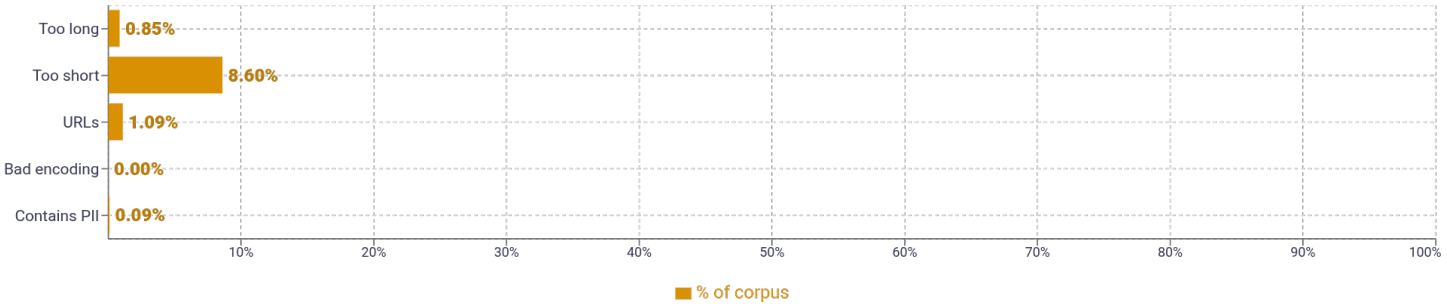
## Frequent n-grams

| SIZE | N-GRAMS | |
|---|---|---|
| 1 | गई \| 34,305,856   लेकिन \| 34,034,332   रही \| 32,714,870   पहले \| 32,086,004   रूप \| 31,435,687 | ⧉ |
| 2 | दी गई \| 4,979,599   बता दें \| 4,911,200   किए गए \| 4,564,097   सोशल मीडिया \| 4,551,281   in hindi \| 4,195,761 | ⧉ |
| 3 | कम से कम \| 2,268,850   मौत हो गई \| 1,711,697   प्रधानमंत्री नरेंद्र मोदी \| 1,603,645   योजना के तहत \| 1,437,433   बारे में जानकारी \| 966,431 | ⧉ |
| 4 | अस्पताल में भर्ती कराया \| 364,708   रूप में जाना जाता \| 346,078   सोशल मीडिया पर वायरल \| 332,837   कम करने में मदद \| 293,845   नाम से जाना जाता \| 292,317 | ⧉ |
| 5 | लोगों की मौत हो गई \| 339,561   नाम से भी जाना जाता \| 212,704   जानकारी के लिए बता दें \| 167,090   जान से मारने की धमकी \| 158,412   view this post on instagram \| 155,315 | ⧉ |

## 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 | | | | |