# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-asm_Beng | 9/16/2025 | Assamese (as) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 446,306 | 6,543,402 | 5,513,688 (84.26 %) | 206M | 1,140,988,880 | 2.79 GB |

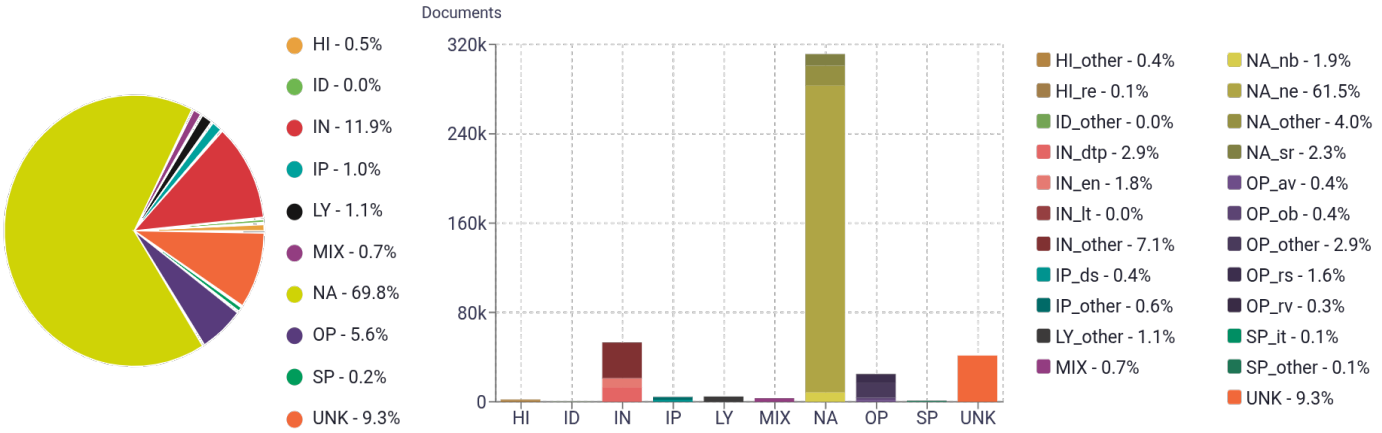
## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| nenow.in | 51K | 11.41% |
| news18.com | 33K | 7.30% |
| asomiyapratidin.in | 31K | 7.00% |
| wikisource.org | 15K | 3.34% |
| eastmojo.com | 13K | 2.97% |
| janambhumi.in | 11K | 2.45% |
| sentinelassam.com | 10K | 2.32% |
| etvbharat.com | 9K | 2.01% |
| dainikagradoot.in | 8.6K | 1.93% |
| nefocus.com | 8.3K | 1.87% |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| com | 236K | 52.84% |
| in | 161K | 36.12% |
| org | 36K | 8.02% |
| gov.in | 2.9K | 0.65% |
| co.in | 1.9K | 0.42% |
| org.in | 942 | 0.21% |
| news | 890 | 0.20% |
| net | 819 | 0.18% |
| com.br | 600 | 0.13% |
| blog | 593 | 0.13% |

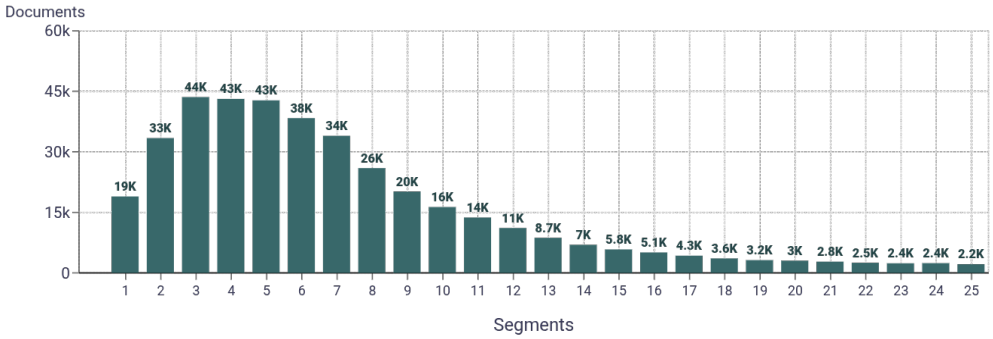
## Register labels




**MT**:3.1% | 14K Documents

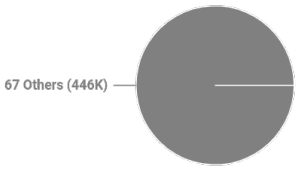
## Documents size (in segments)

**≤ 25** segments **88.38%** (394K documents)
**> 25** segments **11.62%** (52K documents)



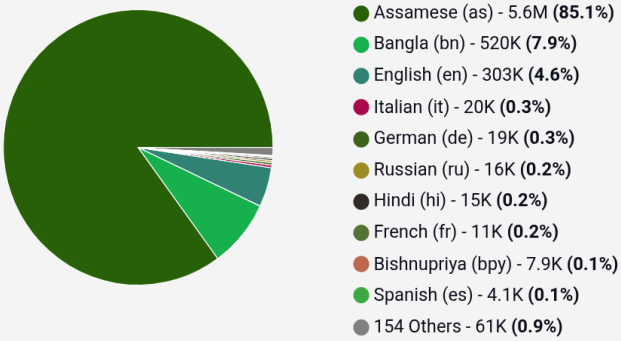

## Document collections

**CC = 99.32%**
**IA = 0.68%**

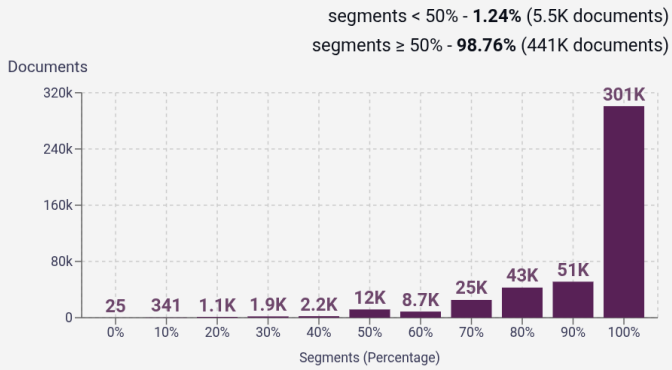


## Language Distribution

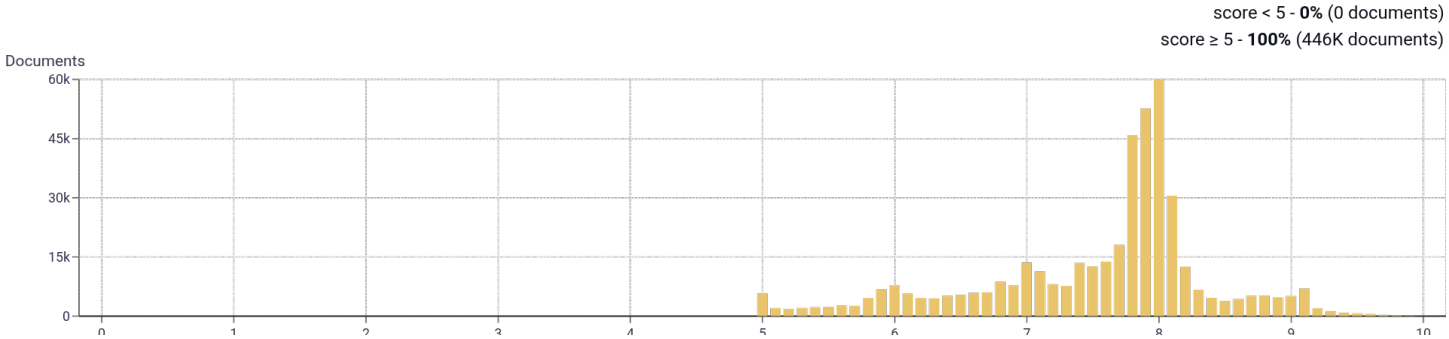
### Number of segments in the Assamese (as) corpus



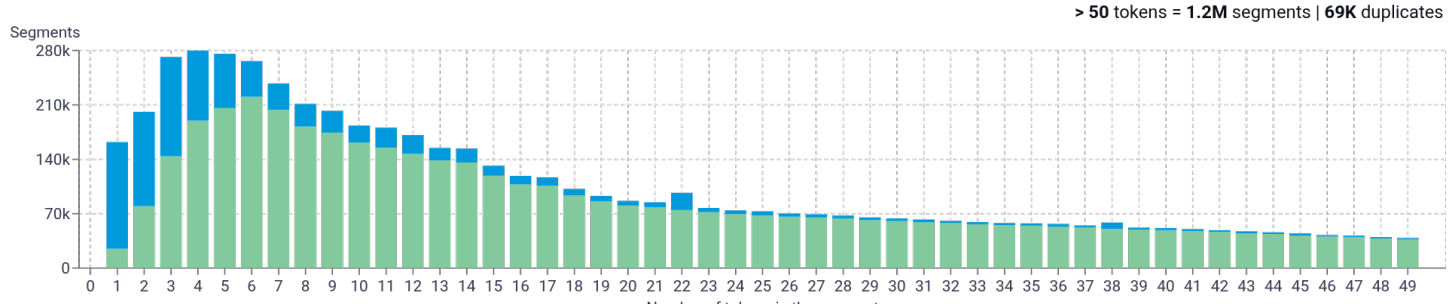

### Percentage of segments in Assamese (as) inside documents



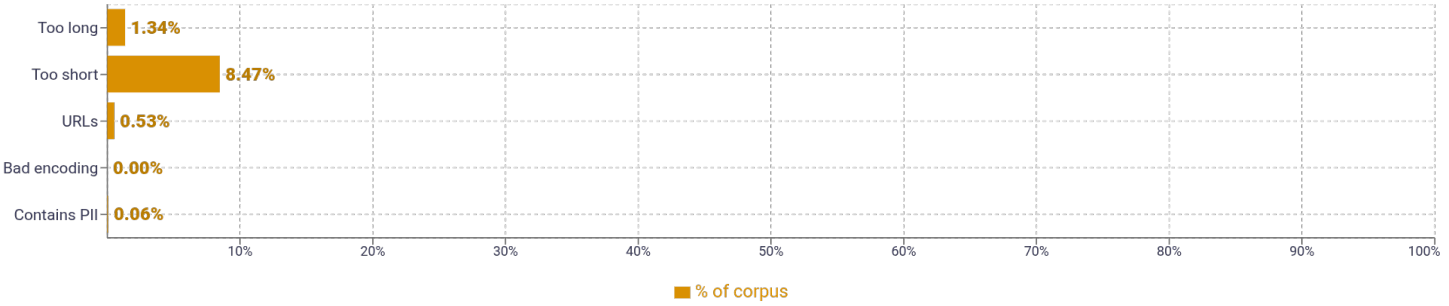

## Distribution of documents by document score




## Segment length distribution by token




## Segment noise distribution



### Frequent n-grams

| SIZE | N-GRAMS |
|---|---|
| 1 | কৰা \| 1,356,189 · হ \| 1,133,037 · কৰি \| 1,122,834 · কৰে \| 1,002,592 · কৰিব \| 862,173 |
| 2 | কৰা হৈছে \| 184,625 · কৰিব পাৰে \| 168,302 · ব পাৰে \| 137,430 · কৰা হ \| 81,712 · কৰা হৈছিল \| 72,612 |
| 3 | জানিব পৰা গৈছে \| 22,348 · হিমন্ত বিশ্ব শৰ্মাই \| 20,531 · ব্রেকিং নিউজ সৰ্বপ্ৰথম \| 16,669 · সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া \| 16,633 · লাইভ নিউজ আপডেট \| 16,633 |
| 4 | বুলি জানিব পৰা গৈছে \| 18,403 · সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ \| 16,633 · বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট \| 16,633 · ব্রেকিং নিউজ সৰ্বপ্ৰথম news18 \| 16,617 · অসমীয়া নিউজ ৱেবছাইট news18 \| 16,617 |
| 5 | সবাতোকৈ বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট \| 16,633 · বিশ্বাসযোগ্য অসমীয়া নিউজ ৱেবছাইট news18 \| 16,617 · অসমীয়াত ব্রেকিং নিউজ সৰ্বপ্ৰথম news18 \| 16,047 · ব্রেকিং নিউজ সৰ্বপ্ৰথম news18 অসমীয়াত \| 14,036 · অসমীয়া নিউজ ৱেবছাইট news18 অসমীয়া \| 14,036 |

## 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 | | | | |