HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-asm_Beng	9/16/2025	Assamese

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

- HI - 0.5%
- ID - 0.0%
- IN - 11.9%
- IP - 1.0%
- LY - 1.1%
- MIX - 0.7%
- NA - 69.8%
- OP - 5.6%
- SP - 0.2%
- UNK - 9.3%

- HI_other - 0.4%
- HI_re - 0.1%
- ID_other - 0.0%
- IN_dtp - 2.9%
- IN_en - 1.8%
- IN_lt - 0.0%
- IN_other - 7.1%
- IP_ds - 0.4%
- IP_other - 0.6%
- LY_other - 1.1%
- MIX - 0.7%
- NA_nb - 1.9%
- NA_ne - 61.5%
- NA_other - 4.0%
- NA_sr - 2.3%
- OP_av - 0.4%
- OP_ob - 0.4%
- OP_other - 2.9%
- OP_rs - 1.6%
- OP_rv - 0.3%
- SP_it - 0.1%
- SP_other - 0.1%
- UNK - 9.3%
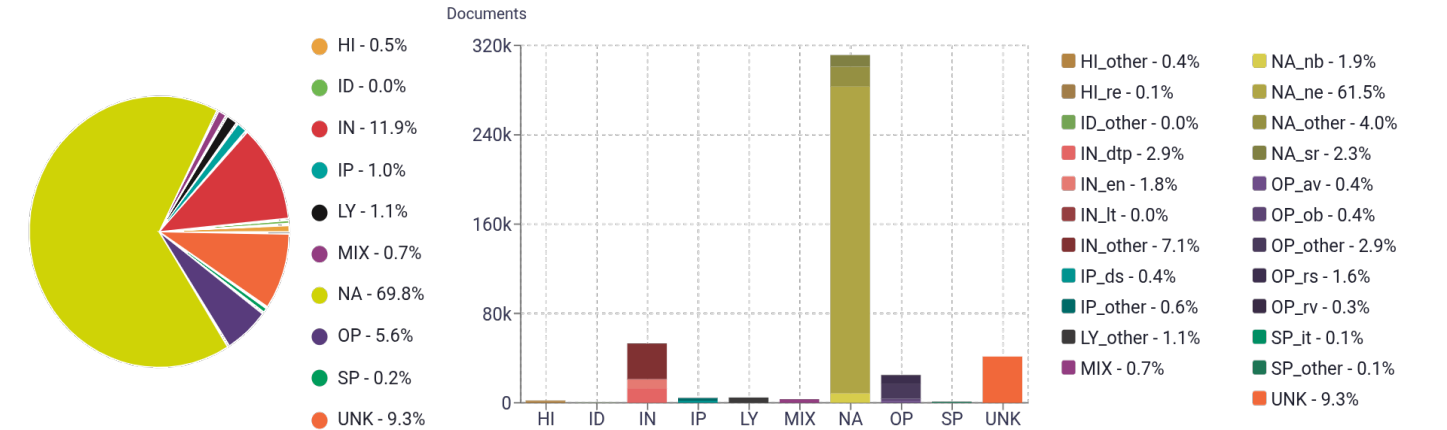
MT:3.1% | 14K Documents

Documents size (in segments) ⓘ

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


Document collections

CC = 99.32%
IA = 0.68%

67 Others (446K)
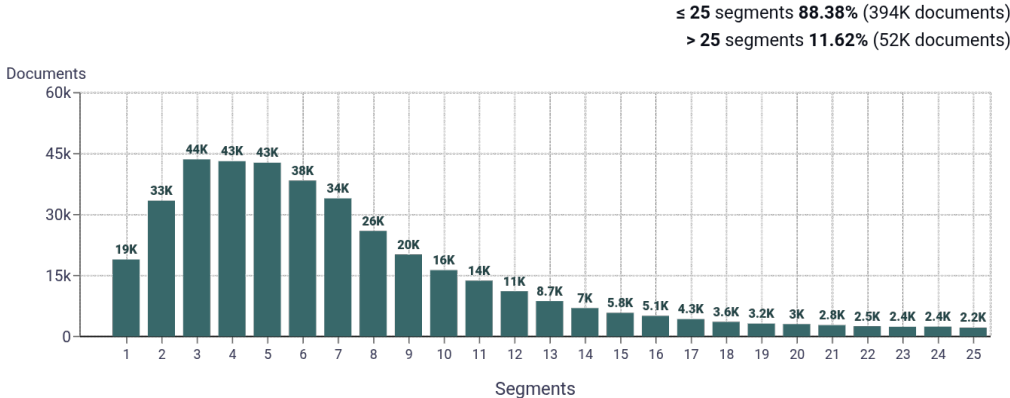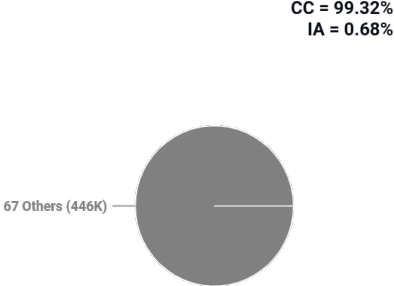
Language Distribution

Number of segments in the Assamese corpus

- Assamese - 5.6M **(85.1%)**
- Bangla - 520K **(7.9%)**
- English - 303K **(4.6%)**
- Italian - 20K **(0.3%)**
- German - 19K **(0.3%)**
- Russian - 16K **(0.2%)**
- Hindi - 15K **(0.2%)**
- French - 11K **(0.2%)**
- Bishnupriya - 7.9K **(0.1%)**
- Spanish - 4.1K **(0.1%)**
- 154 Others - 61K **(0.9%)**
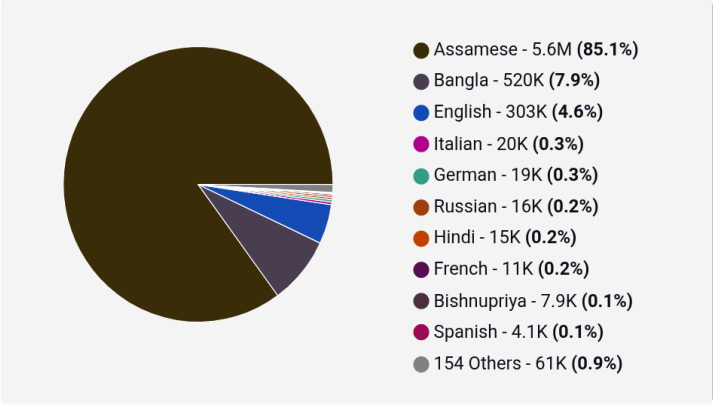
Percentage of segments in Assamese inside documents
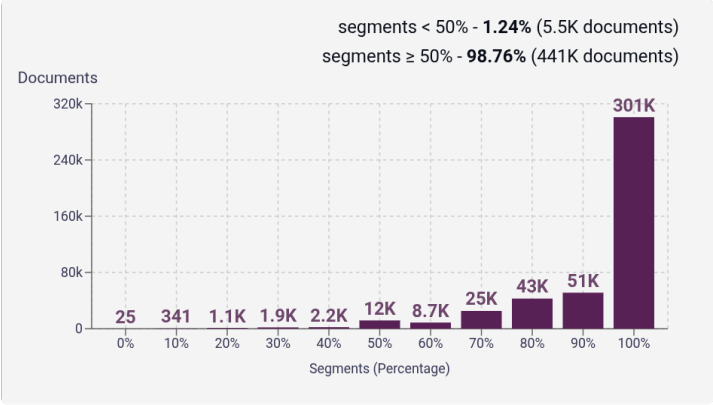
segments < 50% - **1.24%** (5.5K documents)
segments ≥ 50% - **98.76%** (441K documents)

Documents

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
25	341	1.1K	1.9K	2.2K	12K	8.7K	25K	43K	51K	301K

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (446K documents)

Documents

Segment length distribution by token

≤ **49** tokens = **5.3M** segments | **963K** duplicates

> **50** tokens = **1.2M** segments | **69K** duplicates

Segments

Segment noise distribution

- Too long — **1.34%**
- Too short — **8.47%**
- URLs — **0.53%**
- Bad encoding — **0.00%**
- Contains PII — **0.06%**

■ % of corpus
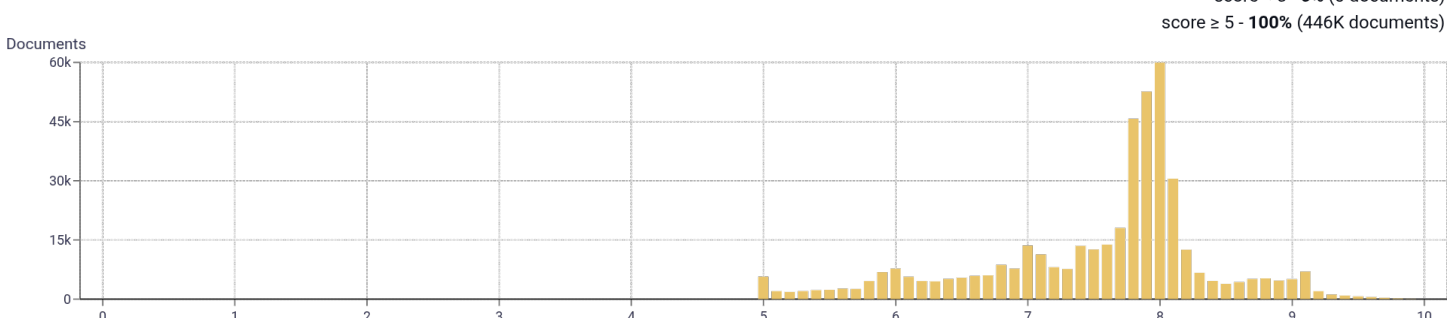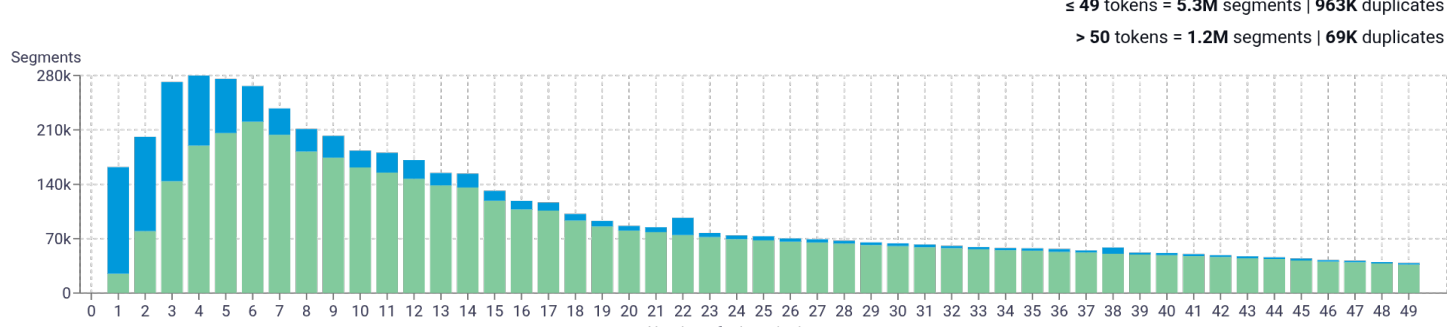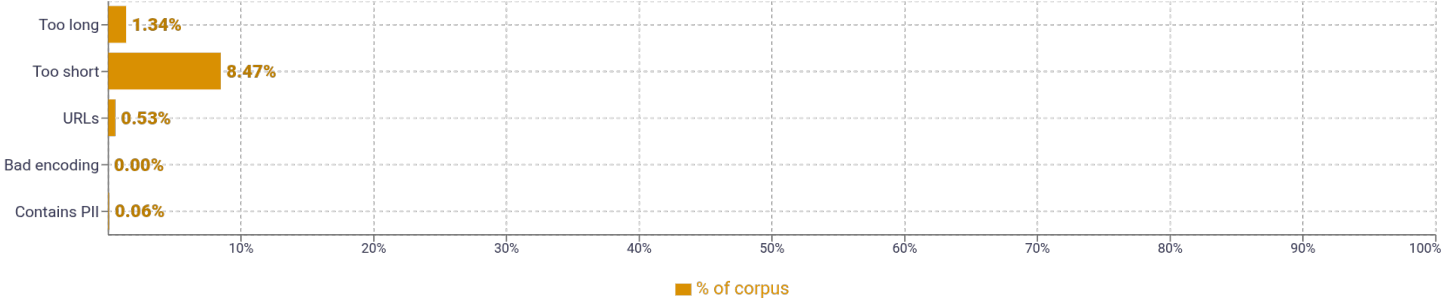
Frequent n-grams

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 ব্ৰেকিং 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 ব্ৰেকিং 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 ব্ৰেকিং নিউজ সৰ্বপ্ৰথম news18 | 16,047 ব্ৰেকিং 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				