HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-mr.tsv	1/21/2025	English (en)	Marathi (mr)

Volumes

Segments	SL tokens	SL characters	SL size
656,962	18M	93,273,281	89.44 MB

TL tokens	TL characters	TL size
18M	101,729,725	251.36 MB

Dataset top 10 domains

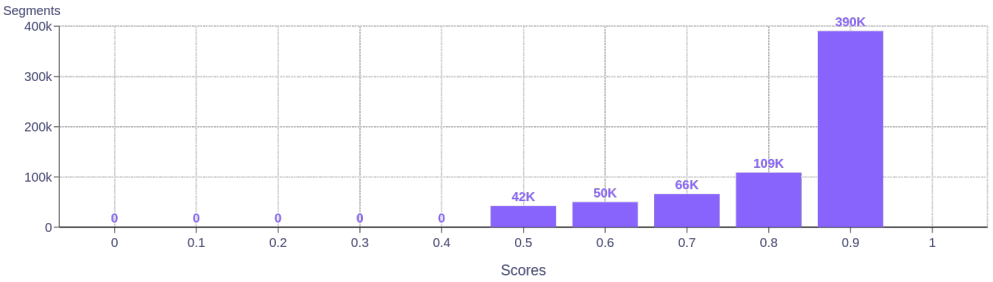
SL domain	Segments	TL domain	Segments
biblegateway.com	12.5%	biblegateway.com	11.1%
educationbro.com	4.2%	wikipedia.org	2.1%
wikipedia.org	2.2%	firstcry.com	1.9%
infeipet.com	1.9%	educationbro.com	1.9%
firstcry.com	1.8%	infeipet.com	1.8%
adda247.com	1.6%	adda247.com	1.8%
whatsapp.com	1.3%	uniquenewsonline.com	1.2%
phoneky.com	1.2%	news18.com	1.2%
uber.com	1.1%	whatsapp.com	1.2%
uniquenewsonline.com	1.1%	wordproject.org	1.0%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	100.0%	com	79.9%
org	10.3%	in	11.9%
in	8.8%	org	9.4%
net	3.5%	net	2.5%
gov.in	1.4%	gov.in	1.3%
co.uk	1.1%	top	0.9%
top	1.0%	co.in	0.8%
co.in	0.9%	info	0.5%
plus	0.8%	plus	0.5%
info	0.7%	zone	0.4%

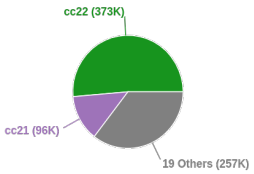
Translation likelihood

≥ 5 = 657K segments | **100.0%**
≥ 8 = 499K segments | **76.0%**
< 5 = 0 segments | **0.0%**



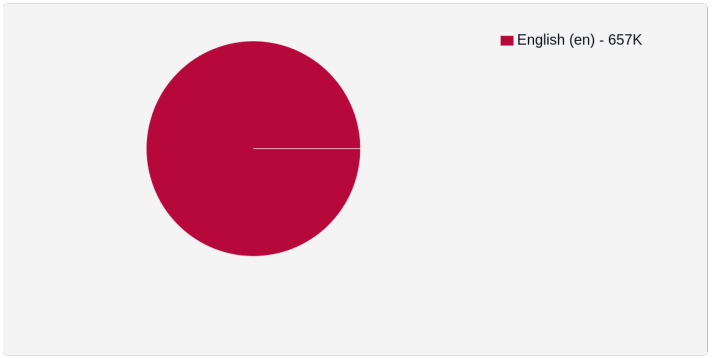
Collections

CC = 80.02%
IA = 19.98%

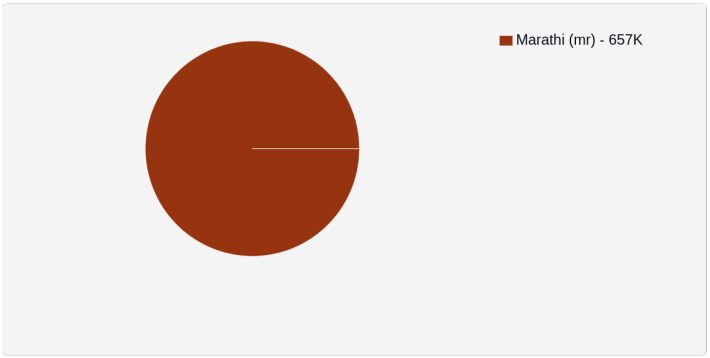


Language Distribution

Source

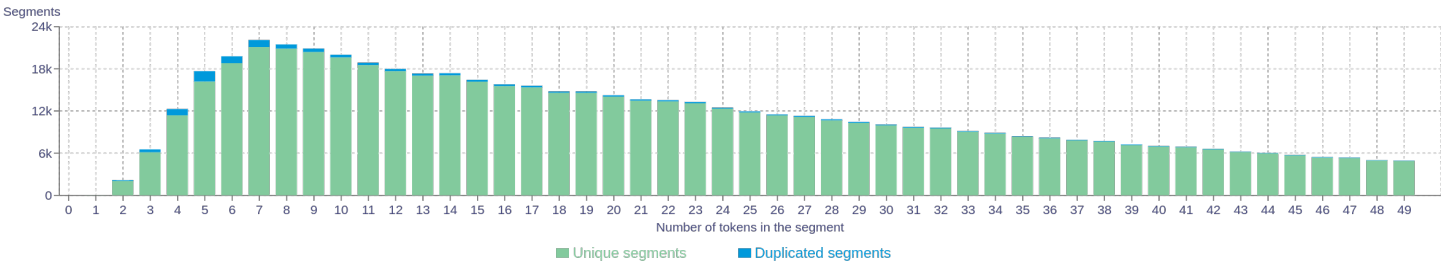


Target



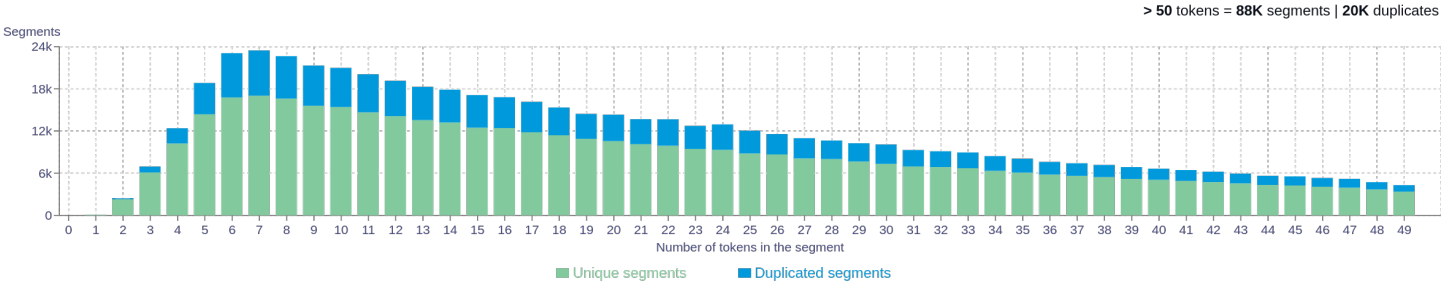
Source segment length distribution by token

<= 49 tokens = **550K** segments | **12K** duplicates
> 50 tokens = **95K** segments | **910** duplicates

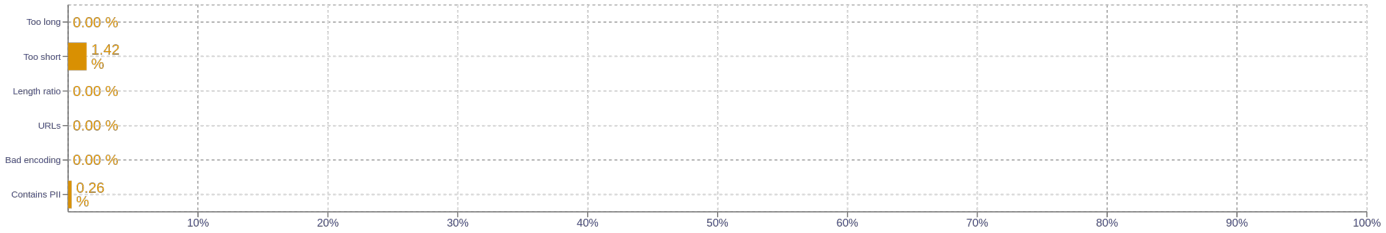


Target segment length distribution by token

<= 49 tokens = **425K** segments | **144K** duplicates
> 50 tokens = **88K** segments | **20K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 35060 · said | 33431 · one | 32555 · india | 29041 · new | 26192
2	happy birthday | 12218 · prime minister | 4363 · new delhi | 4326 · chief minister | 3838 · personal information | 3090
3	jammu and kashmir | 1702 · t20 world cup | 1688 · bank of india | 1666 · petrol and diesel | 1659 · india meteorological department | 1427
4	prime minister narendra modi | 1271 · one of the best | 1255 · reserve bank of india | 994 · heavy to very heavy | 788 · visit the official website | 663
5	board of control for cricket | 774 · control for cricket in india | 766 · heavy to very heavy rainfall | 558 · central board of secondary education | 556 · maharashtra state board of secondary | 480

Target n-grams

Size	n-grams
1	किंवा | 73261 · आपण | 59855 · करण्यासाठी | 45969 · आम्ही | 44597 · तुम्ही | 44500
2	करू शकता | 15004 · हार्दिक शुभेच्छा | 7270 · क्लिक करा | 6250 · जाऊ शकते | 5986 · वाढदिवसाच्या हार्दिक | 5605
3	वाढदिवसाच्या हार्दिक शुभेच्छा | 5397 · खूप खूप शुभेच्छा | 1885 · पंतप्रधान नरेंद्र मोदी | 1802 · साजरा केला जातो | 1497 · वर क्लिक करा | 1291
4	वाढदिवसाच्या खूप खूप शुभेच्छा | 1241 · अगोदर निर्देश केलेल्या बाबीसंबंधी | 851 · निर्देश केलेल्या बाबीसंबंधी बोलताना | 837 · पंतप्रधान नरेंद्र मोदी यांनी | 642 · birthday wishes in marathi | 461
5	अगोदर निर्देश केलेल्या बाबीसंबंधी बोलताना | 832 · निश्चितपणे त्याच्या आकर्षक वैशिष्ट्यांचा आनंद | 375 · आपण निश्चितपणे त्याच्या आकर्षक वैशिष्ट्यांचा | 375 · राज्य माध्यमिक आणि उच्च माध्यमिक | 275 · महाराष्ट्र राज्य माध्यमिक आणि उच्च | 275

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