HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
547,168	15M	74,793,487	71.67 MB

TL tokens	TL characters	TL size
12M	91,352,025	234.6 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
wikipedia.org	10.1%	wikipedia.org	9.1%
bajajfinserv.in	5.3%	bajajfinserv.in	4.7%
educationbro.com	4.0%	wikisource.org	3.8%
onworks.net	3.3%	indianexpress.com	2.1%
vsaduidoma.com	1.8%	educationbro.com	1.8%
adda247.com	1.6%	vsaduidoma.com	1.8%
indianexpress.com	1.6%	adda247.com	1.8%
kurangah.com	1.4%	kurangah.com	1.5%
catholicgallery.org	1.3%	onworks.net	1.3%
jw.org	1.3%	wordplanet.org	1.2%

Dataset top 10 TLDs

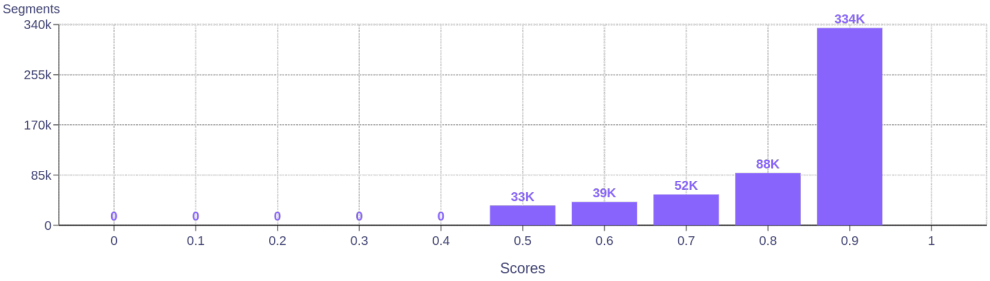
SL domain	Segments	TL domain	Segments
com	80.2%	com	66.2%
org	24.8%	org	23.9%
in	9.5%	in	9.2%
net	7.7%	net	3.9%
info	1.6%	rehab	1.2%
rehab	1.3%	top	0.8%
top	0.8%	zone	0.7%
zone	0.8%	gov.in	0.7%
plus	0.7%	info	0.5%
gov.in	0.7%	plus	0.5%

Translation likelihood

≥ 5 = 547K segments | **100.0%**

≥ 8 = 423K segments | **77.2%**

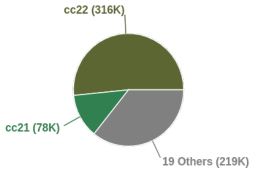
< 5 = 0 segments | **0.0%**



Collections

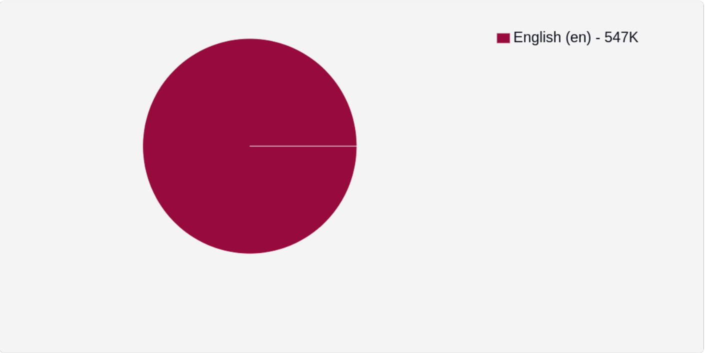
CC = 78.99%

IA = 21.01%

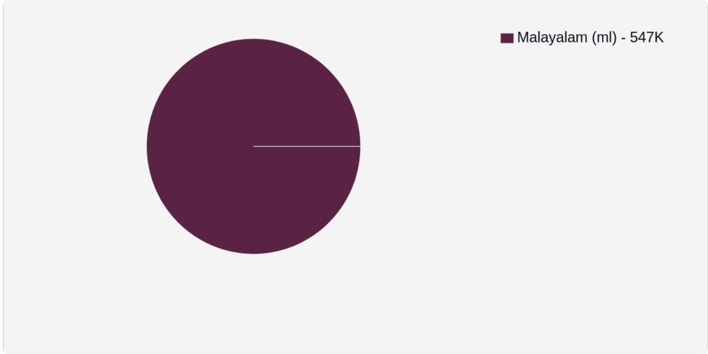


Language Distribution

Source



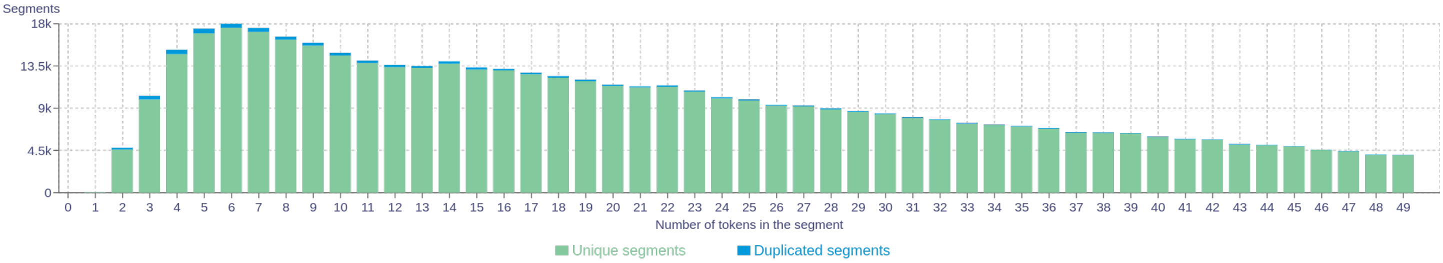
Target



Source segment length distribution by token

<= 49 tokens = **466K** segments | **7.4K** duplicates

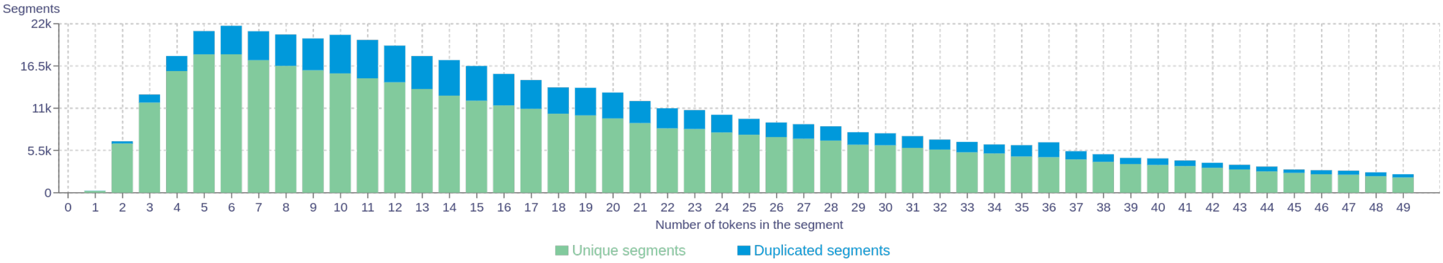
> 50 tokens = **74K** segments | **664** duplicates



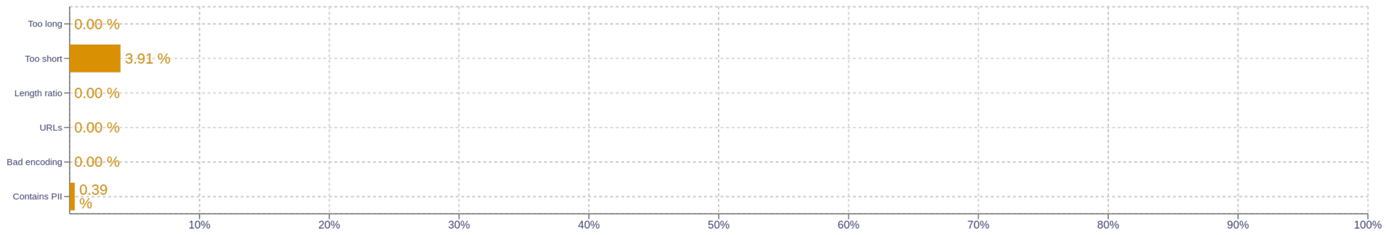
Target segment length distribution by token

<= 49 tokens = **401K** segments | **107K** duplicates

> 50 tokens = **40K** segments | **6.5K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 26606 said | 25901 one | 24533 india | 22051 new | 20968
2	prime minister | 5778 personal loan | 4118 new delhi | 3936 bajaj finserv | 3803 chief minister | 3784
3	jammu and kashmir | 1892 minister narendra modi | 1745 prime minister narendra | 1731 bin rashid al | 1367 mohammed bin rashid | 1357
4	prime minister narendra modi | 1722 archived from the original | 1579 application that can also | 1351 bin rashid al maktoum | 1348 mohammed bin rashid al | 1283
5	also be fetched from https | 1351 mohammed bin rashid al maktoum | 1281 sheikh mohammed bin rashid al | 1019 run online this app named | 981 latest release can be downloaded | 857

Target n-grams

Size	n-grams
1	ഞങ്ങളുടെ | 24000 പുതിയ | 18995 പറഞ്ഞു | 18666 ചെയ്യുക | 13499 അവൻ | 13288
2	ക്ലിക്ക് ചെയ്യുക | 2132 ബജാജ് ഫിൻസെർവ് | 2117 ഹോം ലോൺ | 1626 ആർക്കൈവ് ചെയ്തത് | 1609 ബാങ്ക് ഓഫ് | 1608
3	സൗജന്യ ഓപ്പറേറ്റീവ് സിസ്റ്റങ്ങളിലൊന്നിൽ | 1351 രീതിയിൽ ഓൺലൈനിൽ പ്രവർത്തിപ്പിക്കുന്നതിനായി | 1351 പ്രവർത്തിപ്പിക്കുന്നതിനായി ഇത് onworks | 1351 ഞങ്ങളുടെ സൗജന്യ ഓപ്പറേറ്റീവ് | 1351 എളുപ്പമുള്ള രീതിയിൽ ഓൺലൈനിൽ | 1351
4	സിസ്റ്റങ്ങളിലൊന്നിൽ നിന്ന് ഏറ്റവും എളുപ്പമുള്ള | 1351 ഞങ്ങളുടെ സൗജന്യ ഓപ്പറേറ്റീവ് സിസ്റ്റങ്ങളിലൊന്നിൽ | 1351 ഓൺലൈനിൽ പ്രവർത്തിപ്പിക്കുന്നതിനായി ഇത് onworks | 1351 എളുപ്പമുള്ള രീതിയിൽ ഓൺലൈനിൽ പ്രവർത്തിപ്പിക്കുന്നതിനായി | 1351 സൗജന്യ ഹോസ്റ്റിംഗ് ദാതാവായ onworks | 858
5	സിസ്റ്റങ്ങളിലൊന്നിൽ നിന്ന് ഏറ്റവും എളുപ്പമുള്ള രീതിയിൽ | 1351 രീതിയിൽ ഓൺലൈനിൽ പ്രവർത്തിപ്പിക്കുന്നതിനായി ഇത് onworks | 1351 ഓപ്പറേറ്റീവ് സിസ്റ്റങ്ങളിലൊന്നിൽ നിന്ന് ഏറ്റവും എളുപ്പമുള്ള | 1351 എന്നതിൽ നിന്നും ലഭിക്കാവുന്ന ഒരു ആപ്ലിക്കേഷനാണ് | 984 സൗജന്യമായി ഓൺലൈനായി ഡൗൺലോഡ് ചെയ്ത് പ്രവർത്തിപ്പിക്കുക | 679

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