HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-ta.tsv	1/22/2025	English (en)	Tamil (ta)

Volumes

Segments	SL tokens	SL characters	SL size
1,111,471	30M	151,447,074	145.25 MB

TL tokens	TL characters	TL size
25M	185,182,398	470.48 MB

Dataset top 10 domains

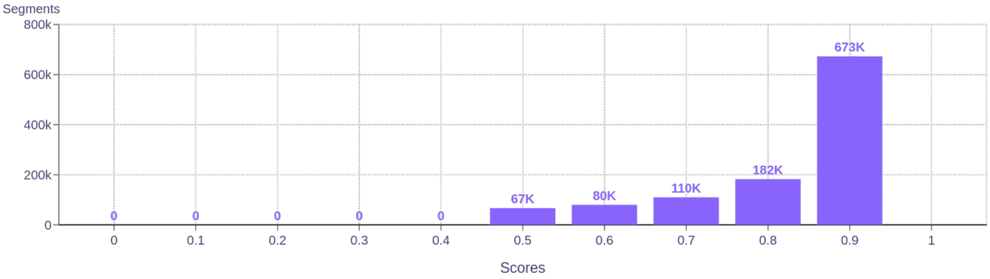
SL domain	Segments	TL domain	Segments
wikipedia.org	10.9%	wikipedia.org	8.7%
downloadastro.com	8.7%	biblegateway.com	7.1%
biblegateway.com	8.0%	downloadastro.com	6.9%
sermoncentral.com	2.5%	wordproject.org	2.9%
bajajfinserv.in	2.4%	bajajfinserv.in	2.1%
educationbro.com	2.3%	websiterating.com	2.1%
websiterating.com	2.1%	wsws.org	1.9%
wsws.org	1.9%	itsmygame.org	1.5%
itsmygame.org	1.7%	catholicgallery.org	1.4%
religion-facts.com	1.6%	indianexpress.com	1.2%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	88.1%	com	70.7%
org	25.6%	org	24.7%
net	7.0%	in	7.7%
in	6.9%	net	5.6%
lk	2.2%	lk	2.4%
info	1.2%	gov.lk	1.0%
gov.lk	1.1%	info	0.7%
plus	0.6%	top	0.6%
top	0.6%	gov.in	0.4%
co.uk	0.5%	edu	0.3%

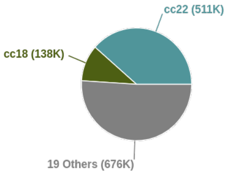
Translation likelihood

≥ 5 = 1.1M segments | **100.0%**
≥ 8 = 855K segments | **76.9%**
< 5 = 0 segments | **0.0%**



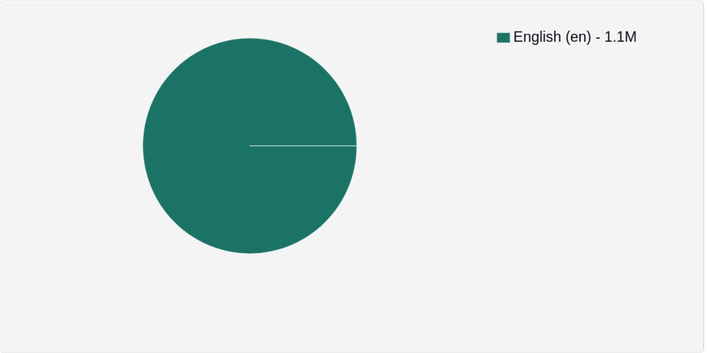
Collections

CC = 68.32%
IA = 31.68%

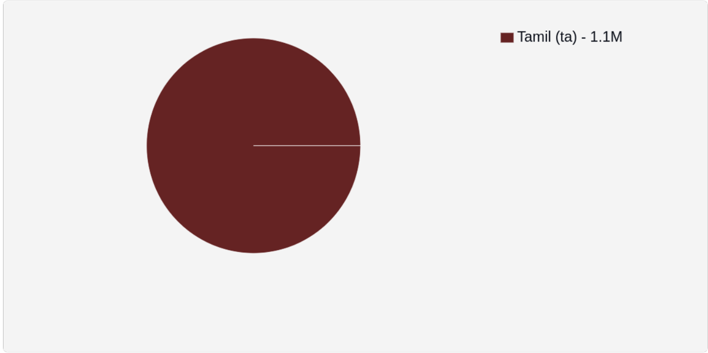


Language Distribution

Source

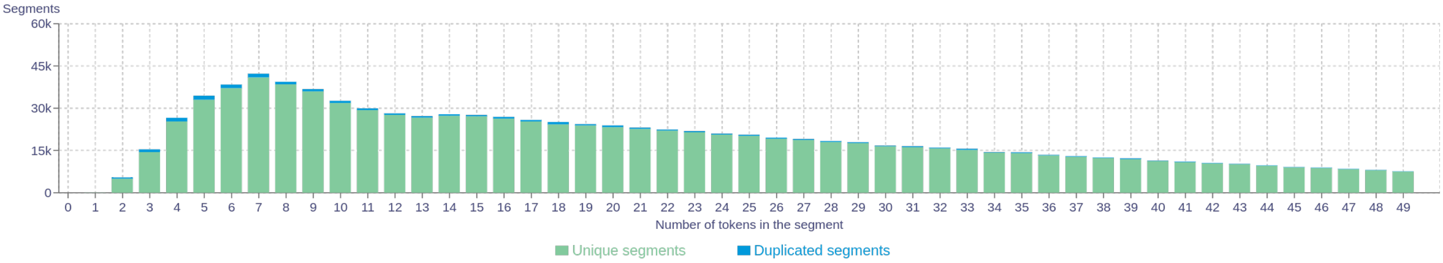


Target



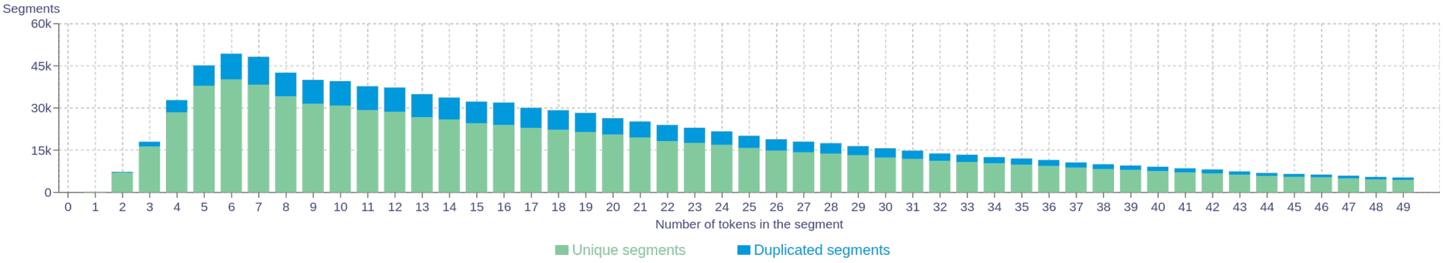
Source segment length distribution by token

<= 49 tokens = **940K** segments | **22K** duplicates
> 50 tokens = **149K** segments | **1.6K** duplicates

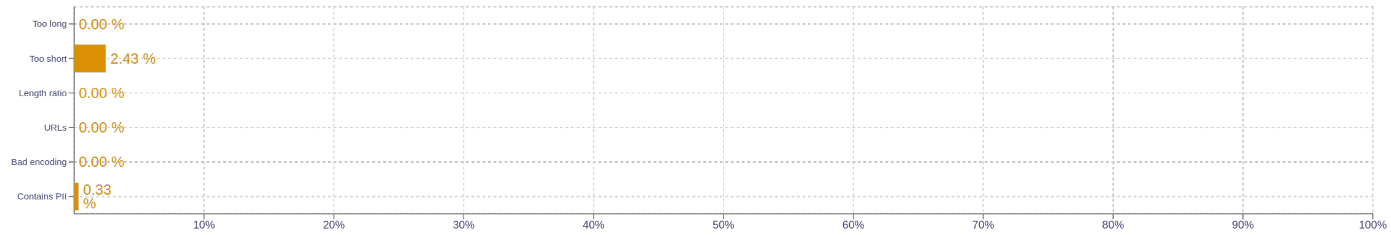


Target segment length distribution by token

<= 49 tokens = **814K** segments | **209K** duplicates
> 50 tokens = **89K** segments | **14K** duplicates



Segment pair noise distribution



Source n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>said | 64218
one | 53527
also | 52852
lord | 52613
people | 43383</td></tr>
<tr><td>2</td><td>sri lanka | 10540
prime minister | 9411
software similar | 6782
tamil nadu | 6440
united states | 5806</td></tr>
<tr><td>3</td><td>happy to recommend | 5029
recommend you programs | 5027
users who downloaded | 4533
number of programs | 2867
minister narendra modi | 2830</td></tr>
<tr><td>4</td><td>recommend you programs like | 5026
prime minister narendra modi | 2770
characteristics of the game | 1832
games like the game | 1627
word of the lord | 1553</td></tr>
<tr><td>5</td><td>happy to recommend you programs | 5027
technical characteristics of the game | 1832
middle east and north africa | 855
latin america and the caribbean | 803
word of the lord came | 609</td></tr>
</table>

Target n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>நீங்கள் | 134614
உங்கள் | 121236
மூலம் | 42038
ஆனால் | 41260
நாங்கள் | 36885</td></tr>
<tr><td>2</td><td>மென்பொருளுக்கு ஒத்த | 7519
ஒத்த மென்பொருட்கள் | 7519
விரும்பிய மென்பொருட்களை | 5069
மென்பொருட்களை பரிந்துரைப்பதில் | 5069
பயனாளிகள் விரும்பிய | 5069</td></tr>
<tr><td>3</td><td>மென்பொருளுக்கு ஒத்த மென்பொருட்கள் | 7519
விரும்பிய மென்பொருட்களை பரிந்துரைப்பதில் | 5069
பயனாளிகள் விரும்பிய மென்பொருட்களை | 5069
மென்பொருட்களை பரிந்துரைப்பதில் மகிழ்கிறோம் | 4828
மென்பொருள்களையும் பதிவிறக்கம் செய்தார்கள் | 4208</td></tr>
<tr><td>4</td><td>பயனாளிகள் விரும்பிய மென்பொருட்களை பரிந்துரைப்பதில் | 5069
விரும்பிய மென்பொருட்களை பரிந்துரைப்பதில் மகிழ்கிறோம் | 4828
மென்பொருளைப் பதிவிறக்கம் செய்த பயனாளிகள் | 4208
மத்திய கிழக்கு மற்றும் வட | 1583
கிழக்கு மற்றும் வட ஆப்பிரிக்கா | 1573</td></tr>
<tr><td>5</td><td>பயனாளிகள் விரும்பிய மென்பொருட்களை பரிந்துரைப்பதில் மகிழ்கிறோம் | 4828
மத்திய கிழக்கு மற்றும் வட ஆப்பிரிக்கா | 1573
கிழக்கு மற்றும் வட ஆப்பிரிக்கா ஒப்பிடும்போது | 757
எண்ணிக்கையிலான மக்கள் தொகை கொண்ட நாடுகளில் | 711
தொகை கொண்ட நாடுகளில் எந்த நாடு | 703</td></tr>
</table>

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