HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
273,430	7.1M	37,769,045	36.19 MB

TL tokens	TL characters	TL size
6.8M	37,527,716	91.46 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
educationbro.com	8.5%	wsws.org	4.3%
theasianparent.com	5.2%	theasianparent.com	4.2%
wikipedia.org	4.7%	educationbro.com	3.8%
wsws.org	4.5%	wikipedia.org	3.4%
ustraveldocs.com	3.5%	blogspot.com	2.5%
manuals.plus	1.3%	ustraveldocs.com	1.6%
nordfx.com	1.3%	jw.org	1.2%
paypal.com	1.1%	nordfx.com	1.1%
blogspot.com	1.1%	manuals.plus	1.0%
jw.org	1.1%	itstechschool.com	1.0%

Dataset top 10 TLDs

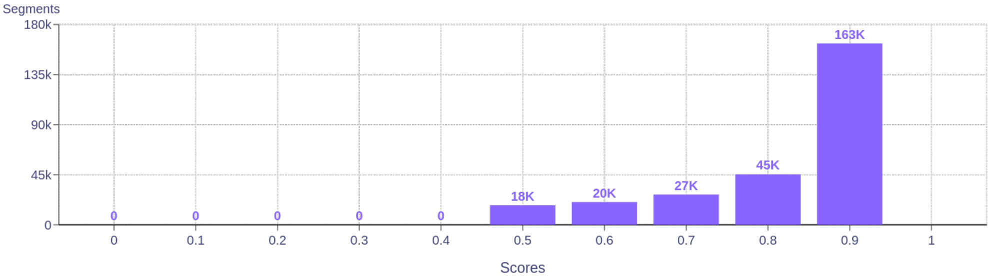
SL domain	Segments	TL domain	Segments
com	91.0%	com	66.3%
org	19.2%	org	15.9%
lk	10.9%	lk	12.8%
gov.lk	6.6%	gov.lk	6.3%
net	4.8%	net	3.7%
plus	1.3%	plus	1.2%
online	0.9%	online	0.9%
zone	0.9%	zone	0.9%
eu	0.9%	it	0.4%
cc	0.8%	eu	0.4%

Translation likelihood

≥ 5 = 273K segments | **100.0%**

≥ 8 = 208K segments | **76.2%**

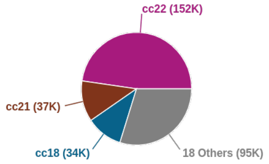
< 5 = 0 segments | **0.0%**



Collections

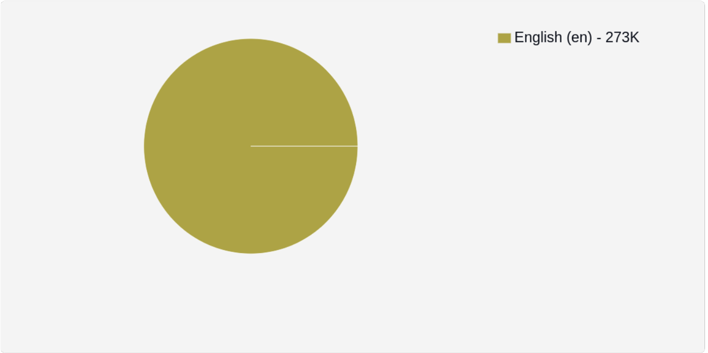
CC = 73.75%

IA = 26.25%

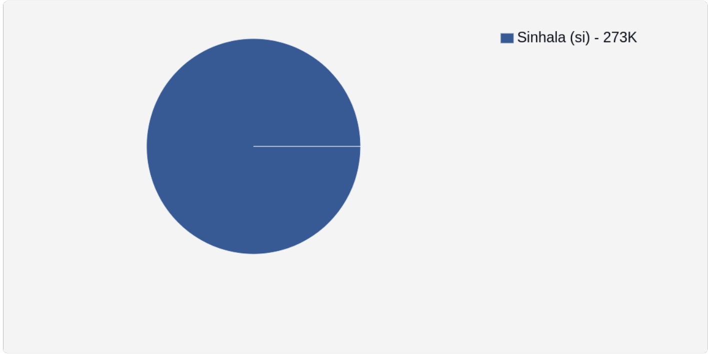


Language Distribution

Source



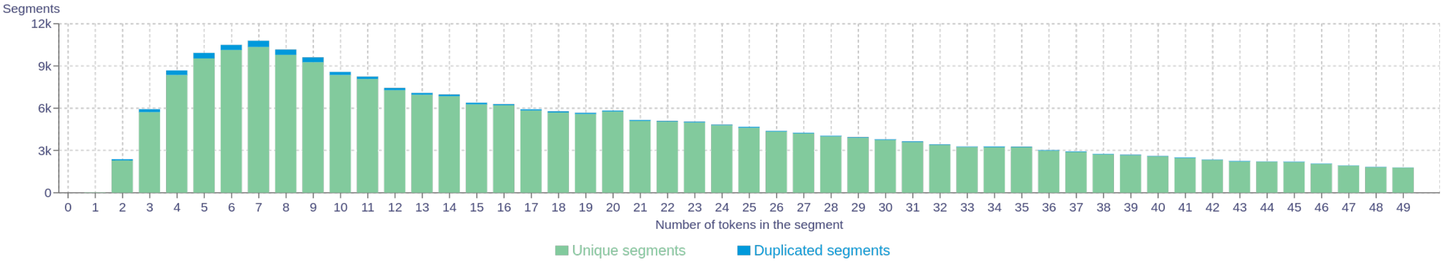
Target



Source segment length distribution by token

<= 49 tokens = **233K** segments | **4.9K** duplicates

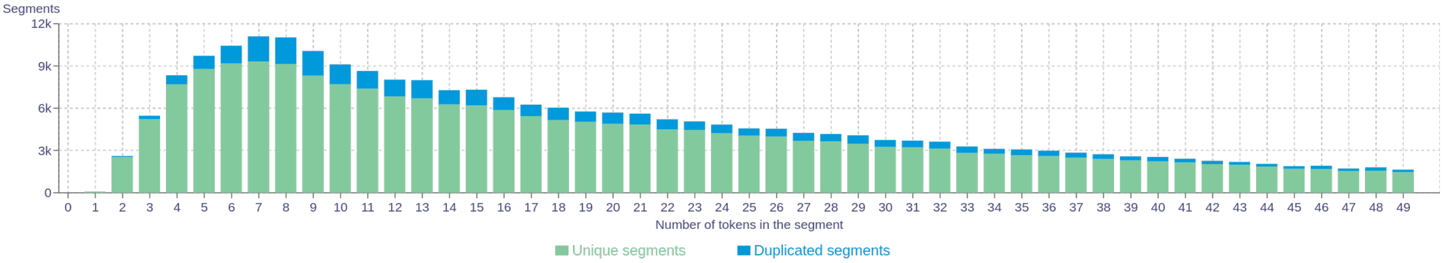
> 50 tokens = **36K** segments | **303** duplicates



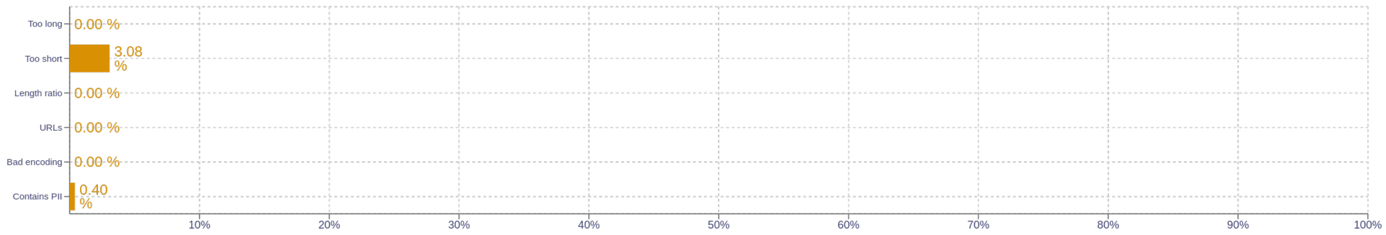
Target segment length distribution by token

<= 49 tokens = **211K** segments | **31K** duplicates

> 50 tokens = **31K** segments | **3.4K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	new | 13229 sri | 12575 also | 11945 information | 11340 one | 10915
2	sri lanka | 10437 new year | 3906 happy new | 3446 united states | 1975 working class | 1780
3	happy new year | 3419 part time jobs | 885 name of jesus | 723 socialist equality party | 668 terms and conditions | 618
4	best part time jobs | 458 commander of the navy | 435 replydelete happy new year | 344 committee of the fourth | 323 north korean leader kim | 320
5	international committee of the fourth | 323 committee of the fourth international | 322 central bank of sri lanka | 265 part time jobs is exclusively | 247 time jobs is exclusively sharing | 234

Target n-grams

Size	n-grams
1	කිරීම | 39128 හෝ | 33303 කිරීමට | 29686 ඔබ | 29605 ඇත | 29452
2	කරන ලද | 6019 කළ හැකිය | 5824 ශ්‍රී ලංකා | 5160 කර ඇත | 4670 කළ හැකි | 4561
3	ලබා ගත හැකි | 1315 භාවිතා කළ හැකිය | 956 ලබා ගත හැකිය | 925 සුබ නව වසරක් | 918 අලුත් අවුරුද්දක් වේවා | 914
4	සුබ නව වසරක් වේවා | 769 සුභ නව වසරක් වේවා | 667 සුභ අලුත් අවුරුද්දක් වේවා | 490 හොඳම අඩ කාල රැකියා | 427 සුබ අලුත් අවුරුද්දක් වේවා | 402
5	මිල ගණන් හා ලබා ගත | 302 ගණන් හා ලබා ගත හැකි | 302 සුභම සුභ නව වසරක් වේවා | 236 part time jobs is exclusively | 223 time jobs is exclusively sharing | 214

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