HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-bn.tsv	1/23/2025	English (en)	Bangla (bn)

Volumes

Segments	SL tokens	SL characters	SL size
2,328,136	60M	311,250,028	298.05 MB

TL tokens	TL characters	TL size
57M	331,391,956	822.86 MB

Dataset top 10 domains

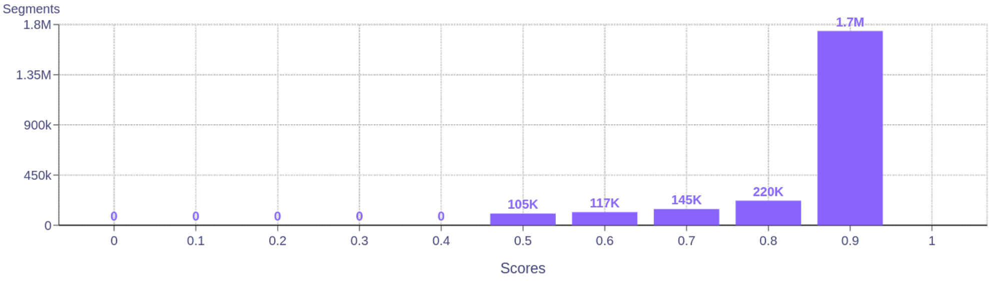
SL domain	Segments	TL domain	Segments
softoware.net	6.0%	wikipedia.org	5.0%
wikipedia.org	5.7%	softoware.net	4.8%
educationbro.com	2.5%	globalvoices.org	1.3%
itsmygame.org	1.7%	khabarsouthasia.com	1.3%
khabarsouthasia.com	1.6%	itsmygame.org	1.3%
globalvoices.org	1.5%	websiterating.com	1.1%
mozilla.org	1.4%	globalvoicesonline.org	1.0%
globalvoicesonline.org	1.2%	educationbro.com	1.0%
androware.net	1.1%	apsva.us	0.9%
websiterating.com	1.1%	androware.net	0.8%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	104.2%	com	76.3%
org	19.4%	org	15.3%
net	13.4%	net	10.6%
in	3.4%	in	2.9%
ru	1.8%	ru	1.7%
gov	1.2%	com.bd	1.3%
us	1.0%	gov	1.0%
co.uk	0.7%	us	0.9%
com.bd	0.7%	info	0.6%
io	0.6%	io	0.5%

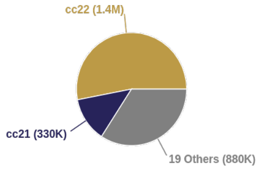
Translation likelihood

≥ 5 = 2.3M segments | **100.0%**
≥ 8 = 2M segments | **84.2%**
< 5 = 0 segments | **0.0%**



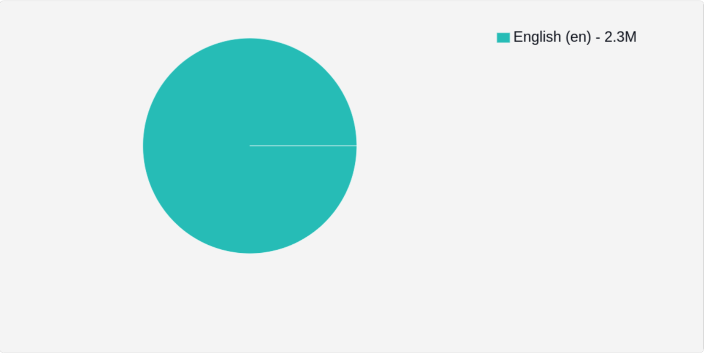
Collections

CC = 78.44%
IA = 21.56%

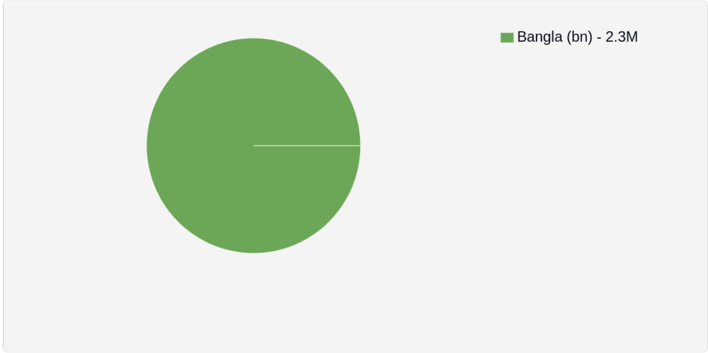


Language Distribution

Source

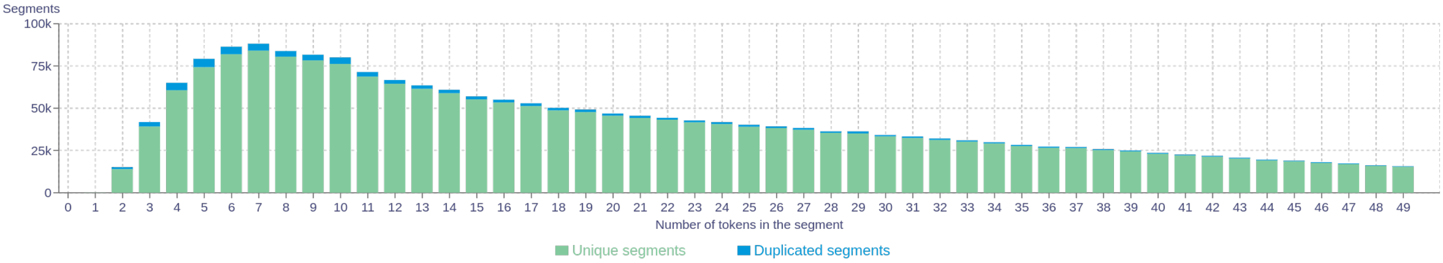


Target



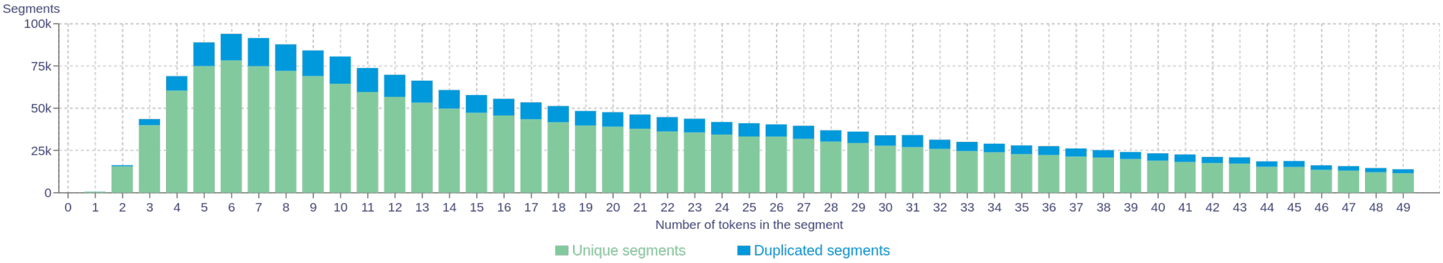
Source segment length distribution by token

<= 49 tokens = **2M** segments | **71K** duplicates
> 50 tokens = **280K** segments | **5.2K** duplicates

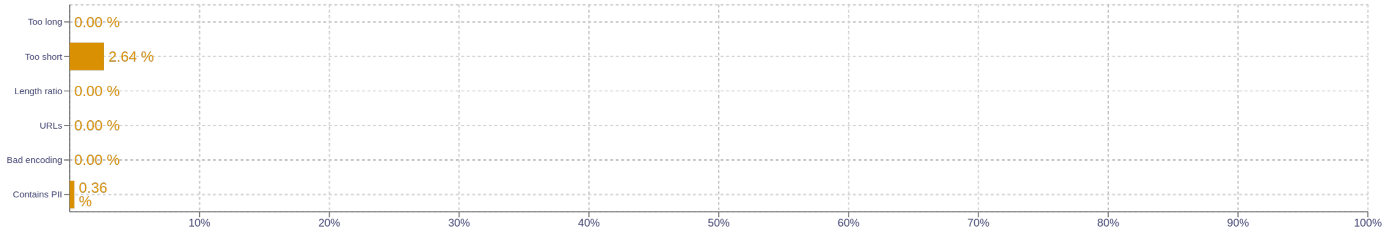


Target segment length distribution by token

<= 49 tokens = **1.7M** segments | **369K** duplicates
> 50 tokens = **241K** segments | **38K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 106646 new | 94076 one | 91558 said | 90677 use | 88011
2	prime minister | 20754 united states | 13697 world cup | 11795 personal information | 9385 privacy policy | 7724
3	peace and blessings | 7312 like the game | 5291 president donald trump | 4821 t20 world cup | 4525 united arab emirates | 4172
4	prime minister sheikh hasina | 3329 link to a friend | 3328 game with the world | 3328 paste in the html | 3317 code of your site | 3317
5	peace and blessings of allaah | 3449 blessings of allaah be upon | 3446 friend or all your friends | 3328 copy and send the link | 3328 copy the code and paste | 3319

Target n-grams

Size	n-grams
1	করুন | 223707 সাথে | 158080 সময় | 100434 এক | 86732 তথ্য | 84089
2	ক্লিক করুন | 26171 সাথে যোগাযোগ | 14021 ব্যক্তিগত তথ্য | 11115 ডাউনলোড করুন | 10627 নির্বাচন করুন | 10152
3	সাথে যোগাযোগ করুন | 5611 প্রেসিডেন্ট ডোনাল্ড ট্রাম্প | 3841 সাকিব আল হাসান | 3834 প্রধানমন্ত্রী শেখ হাসিনা | 3587 মার্কিন প্রেসিডেন্ট ডোনাল্ড | 3559
4	সাইটের html কোড কোড | 3333 কোড কোড এবং পেস্ট | 3333 কোড এবং পেস্ট কপি | 3333 কপি করুন. আপনি খেলা | 3333 বিশ্বের সঙ্গে খেলা শেয়ার | 3330
5	পেস্ট কপি করুন. আপনি খেলা | 3333 কোড কোড এবং পেস্ট কপি | 3333 কোড এবং পেস্ট কপি করুন. | 3333 html কোড কোড এবং পেস্ট | 3333 বিশ্বের সঙ্গে খেলা শেয়ার করুন | 3330

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