HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
1,399,893	40M	207,041,592	198.52 MB

TL tokens	TL characters	TL size
52M	223,909,373	373.21 MB

Dataset top 10 domains

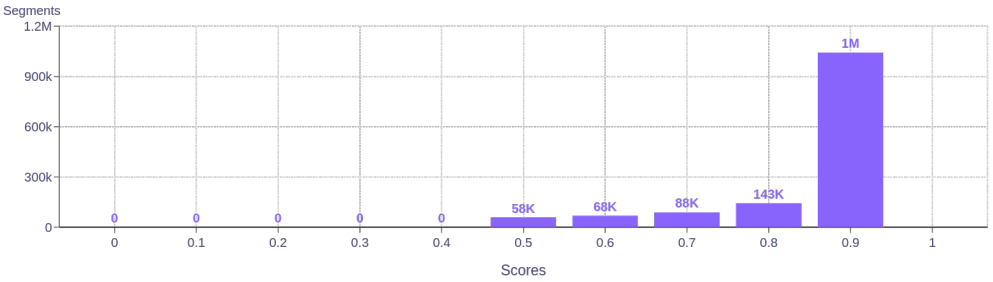
SL domain	Segments	TL domain	Segments
biblegateway.com	5.7%	biblegateway.com	5.1%
khabarsouthasia.com	2.7%	khabarsouthasia.com	2.2%
itsmygame.org	2.7%	kiiky.com	2.1%
kiiky.com	2.1%	itsmygame.org	2.0%
educationbro.com	2.0%	wikipedia.org	1.9%
wikipedia.org	2.0%	websiterating.com	1.7%
websiterating.com	1.7%	rlhymersjr.com	1.6%
rlhymersjr.com	1.7%	asia-news.com	1.6%
vessoft.com	1.7%	arynews.tv	1.5%
asia-news.com	1.7%	flashgames312.com	1.4%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	93.3%	com	73.4%
org	14.8%	org	12.4%
net	8.3%	net	5.6%
tv	2.4%	pk	3.7%
gov	2.1%	tv	3.3%
com.pk	2.0%	com.pk	2.6%
pk	1.8%	gov	1.9%
us	1.7%	us	1.5%
co.uk	1.5%	info	1.4%
info	1.4%	in	0.8%

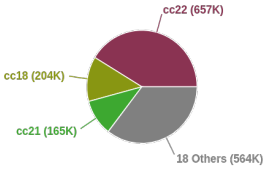
Translation likelihood

≥ 5 = 1.4M segments | **100.0%**
≥ 8 = 1.2M segments | **84.7%**
< 5 = 0 segments | **0.0%**



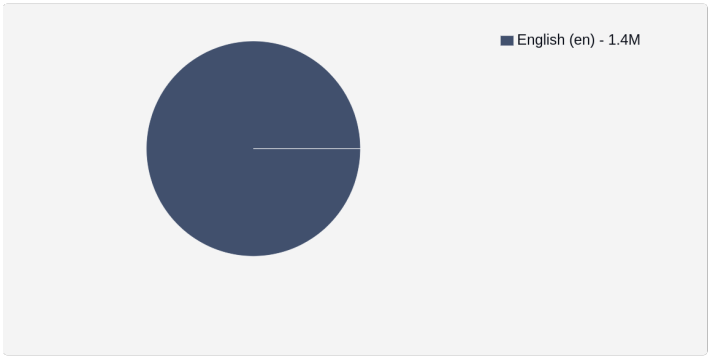
Collections

CC = 70.91%
IA = 29.09%

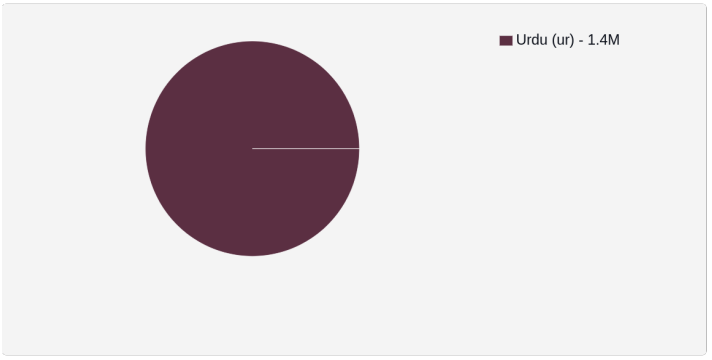


Language Distribution

Source

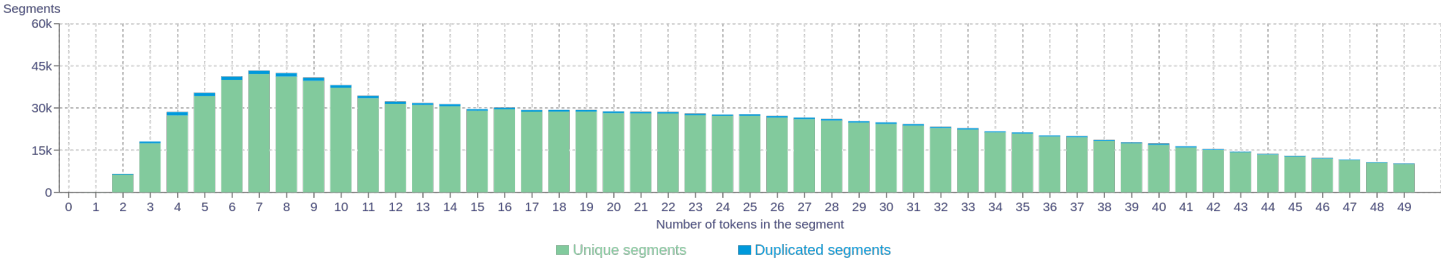


Target



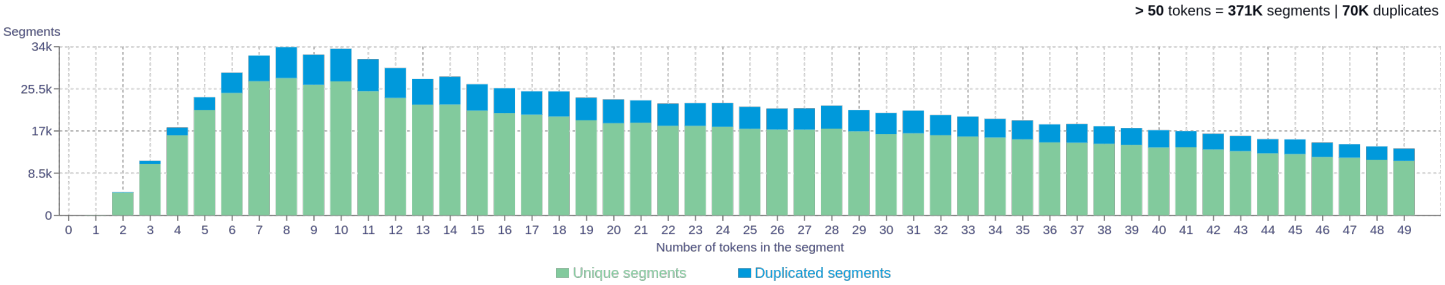
Source segment length distribution by token

<= 49 tokens = **1.2M** segments | **27K** duplicates
> 50 tokens = **203K** segments | **2.7K** duplicates

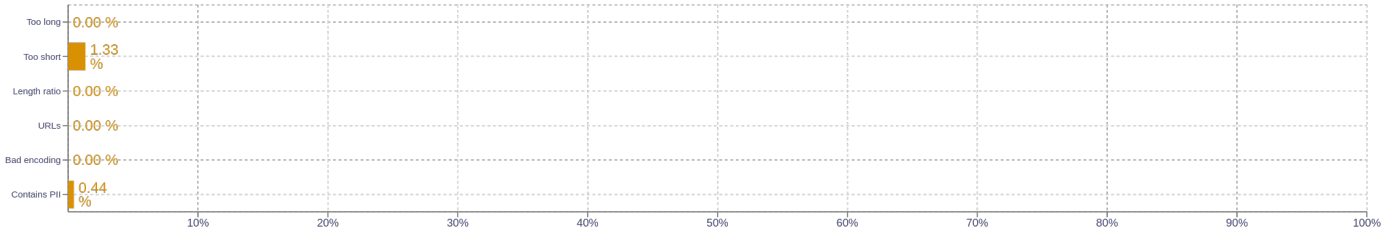


Target segment length distribution by token

<= 49 tokens = **835K** segments | **194K** duplicates
> 50 tokens = **371K** segments | **70K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	said | 103229 · one | 71795 · also | 70212 · people | 59417 · pakistan | 57993
2	prime minister | 16898 · united states | 10994 · personal information | 6486 · imran khan | 6052 · high quality | 5986
3	like the game | 5903 · peace and blessings | 5203 · send the link | 4392 · share the game | 4389 · copy and send | 4388
4	link to a friend | 4390 · game with the world | 4388 · paste in the html | 4375 · code of your site | 4375 · games and download free | 2909
5	friend or all your friends | 4388 · copy and send the link | 4388 · copy the code and paste | 4376 · paste in the html code | 4375 · html code of your site | 4375

Target n-grams

Size	n-grams
1	1231967 | میں · 876049 | سے · 826725 | کو · 596401 | کا · 483963 | نے
2	110441 | آپ کو · 56717 | بارے میں · 54816 | نے کہا · 54806 | کر سکتے · 50524 | سب سے
3	33386 | کرنے کے لیے · 15303 | سب سے زیادہ · 12862 | اس کے علاوہ · 12461 | انہوں نے کہا · 10343 | کیا جا سکتا
4	6784 | ویب سائٹ پر کھیل · 6668 | آپ کی ویب سائٹ · 4395 | آپ کے تمام دوستوں · 4391 | یا آپ کے تمام · 4389 | کو لنک بھیج دیں
5	6779 | اپنی ویب سائٹ پر کھیل · 4391 | آپ کے تمام دوستوں کو · 4389 | یا آپ کے تمام دوستوں · 4389 | کوڈ کی کاپی اور چسپاں · 4389 | کوڈ میں کوڈ کی کاپی

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