HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
902,962	24M	121,140,530	116.04 MB

TL tokens	TL characters	TL size
21M	134,028,027	333.89 MB

Dataset top 10 domains

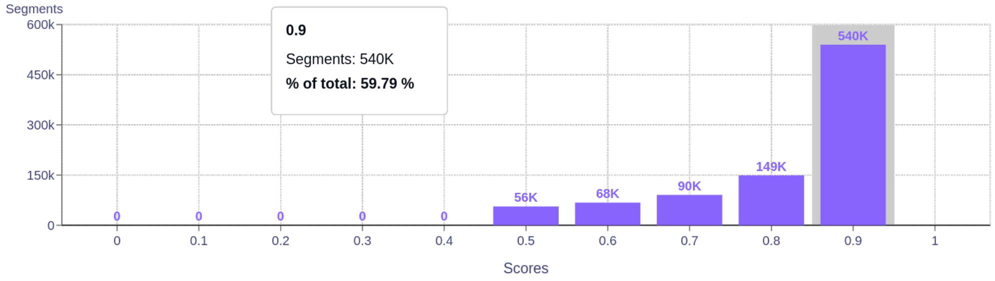
SL domain	Segments	TL domain	Segments
wikipedia.org	26.4%	wikipedia.org	19.1%
itsmygame.org	4.5%	itsmygame.org	3.6%
bajajfinserv.in	3.5%	bajajfinserv.in	3.1%
educationbro.com	3.4%	adda247.com	2.6%
adda247.com	2.3%	blogspot.com	2.3%
vessoft.com	1.7%	educationbro.com	1.7%
boldsky.com	1.6%	boldsky.com	1.6%
biblecloud.com	1.4%	vessoft.in	1.3%
worldwatch.is	1.1%	angelone.in	1.1%
angelone.in	1.1%	richesinchrist.com	0.9%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	80.0%	com	65.2%
org	40.1%	org	29.1%
in	9.5%	in	10.1%
net	6.3%	net	4.7%
gov.in	1.6%	gov.in	1.5%
is	1.1%	top	1.0%
top	1.1%	info	0.7%
info	0.7%	online	0.6%
plus	0.7%	co.in	0.5%
co.in	0.7%	plus	0.5%

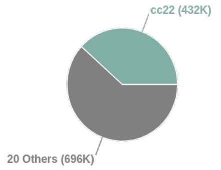
Translation likelihood

≥ 5 = 903K segments | **100.0%**
≥ 8 = 689K segments | **76.3%**
< 5 = 0 segments | **0.0%**



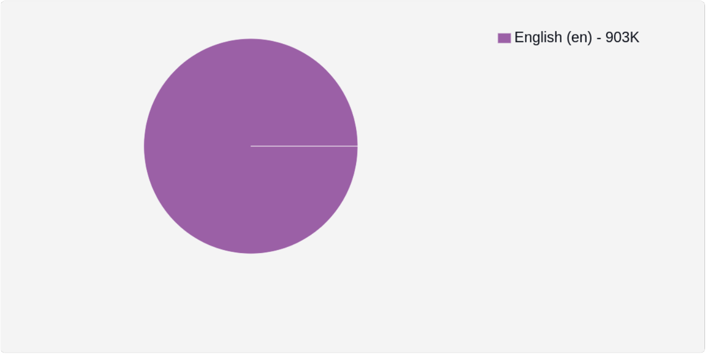
Collections

CC = 69.37%
IA = 30.63%

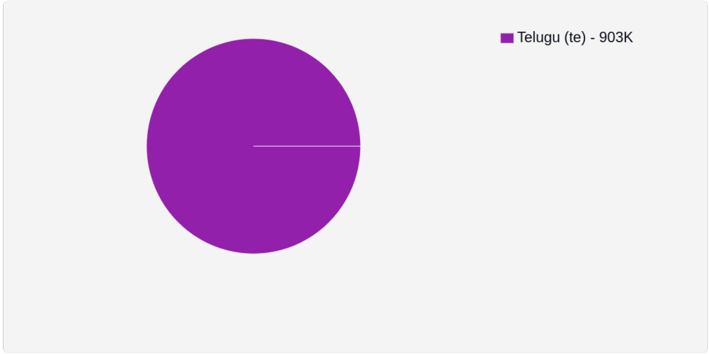


Language Distribution

Source

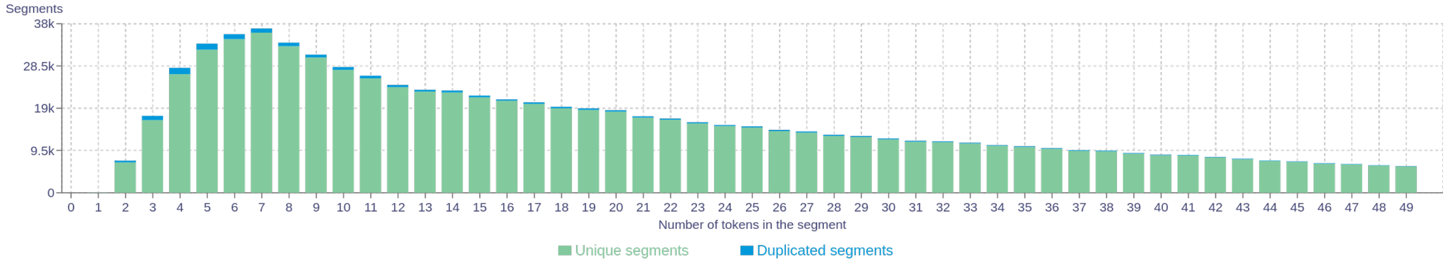


Target



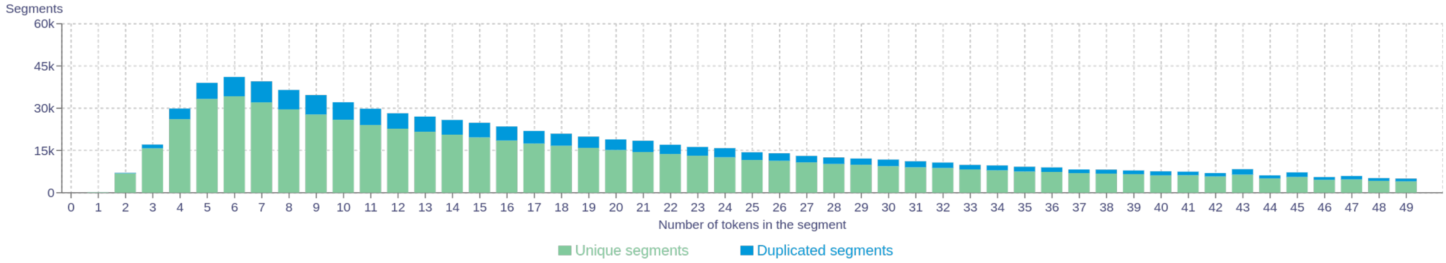
Source segment length distribution by token

<= 49 tokens = **763K** segments | **17K** duplicates
> 50 tokens = **123K** segments | **2.1K** duplicates

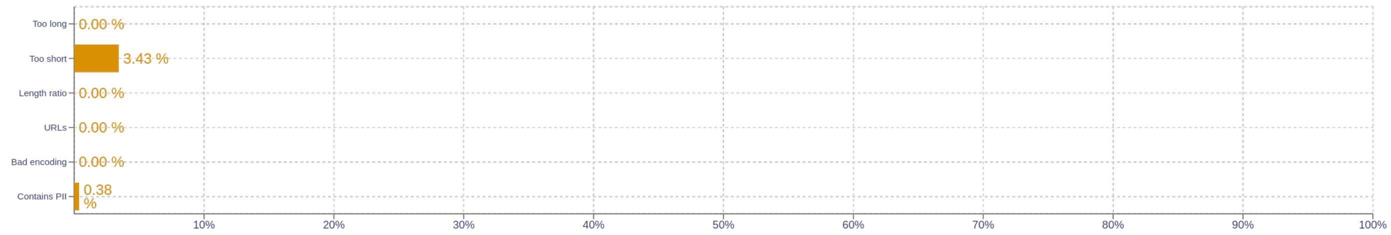


Target segment length distribution by token

<= 49 tokens = **664K** segments | **149K** duplicates
> 50 tokens = **90K** segments | **16K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 47689 game | 39551 one | 39093 new | 34364 use | 30480
2	united states | 4860 html code | 4654 bajaj finserv | 4518 personal information | 4029 prime minister | 3893
3	like the game | 7344 send the link | 4658 share the game | 4657 copy and send | 4656 copy the code | 4633
4	link to a friend | 4656 game with the world | 4656 paste in the html | 4628 code of your site | 4628 games like the game | 2698
5	friend or all your friends | 4656 copy and send the link | 4656 paste in the html code | 4628 html code of your site | 4628 copy the code and paste | 4628

Target n-grams

Size	n-grams
1	ఈ | 152080 మీ | 136083 మీరు | 131368 యొక్క | 129523 లేదా | 100135
2	మీరు మీ | 9932 వెబ్ సైట్ | 9253 సైట్ లో | 8230 మీ వెబ్ | 8024 లో గేమ్ | 7557
3	వెబ్ సైట్ లో | 7973 మీ వెబ్ సైట్ | 7686 సైట్ లో గేమ్ | 7544 మీ సైట్ యొక్క | 4709 లేదా అన్ని మీ | 4662
4	మీ వెబ్ సైట్ లో | 7596 వెబ్ సైట్ లో గేమ్ | 7544 స్నేహితులతో లింక్ పంపడానికి ఉంటే | 4661 స్నేహితుడు లేదా అన్ని మీ | 4661 లేదా అన్ని మీ స్నేహితులతో | 4661
5	మీ వెబ్ సైట్ లో గేమ్ | 7544 స్నేహితుడు లేదా అన్ని మీ స్నేహితులతో | 4661 లేదా అన్ని మీ స్నేహితులతో లింక్ | 4661 మీ స్నేహితులతో లింక్ పంపడానికి ఉంటే | 4661 కాపీ ఇష్టం ఒక స్నేహితుడు లేదా | 4661

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