HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-hi.tsv	1/27/2025	English (en)	Hindi (hi)

Volumes

Segments	SL tokens	SL characters	SL size
9,926,620	250M	1,295,203,875	1.21 GB

TL tokens	TL characters	TL size
300M	1,382,636,071	3.21 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
alibaba.com	19.4%	alibaba.com	13.9%
google.com	7.2%	google.com	2.8%
wikipedia.org	2.3%	wikipedia.org	1.8%
agoda.com	2.2%	wikihow.com	1.5%
microsoft.com	2.1%	microsoft.com	1.4%
wikihow.com	1.6%	affairscloud.com	1.4%
masterstudies.com	1.6%	masterstudies.in	1.3%
affairscloud.com	1.4%	agoda.com	1.1%
biblegateway.com	1.1%	biblegateway.com	0.9%
jagranjosh.com	1.0%	jagranjosh.com	0.9%

Dataset top 10 TLDs

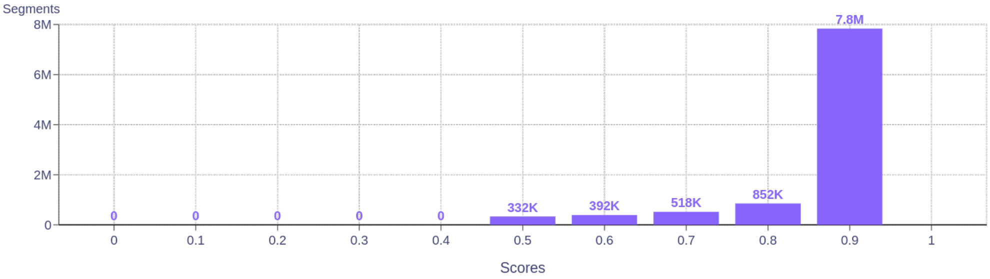
SL domain	Segments	TL domain	Segments
com	118.6%	com	88.4%
org	8.9%	in	12.2%
in	7.8%	org	7.0%
net	2.9%	net	2.2%
gov.in	1.1%	co.in	1.2%
info	0.9%	gov.in	1.0%
co.in	0.9%	info	0.9%
ru	0.9%	ru	0.8%
co.uk	0.7%	gr	0.4%
fm	0.5%	nic.in	0.4%

Translation likelihood

≥ 5 = 9.9M segments | **100.0%**

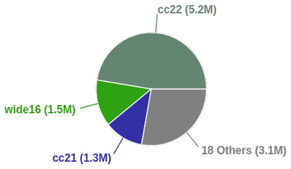
≥ 8 = 8.7M segments | **87.5%**

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



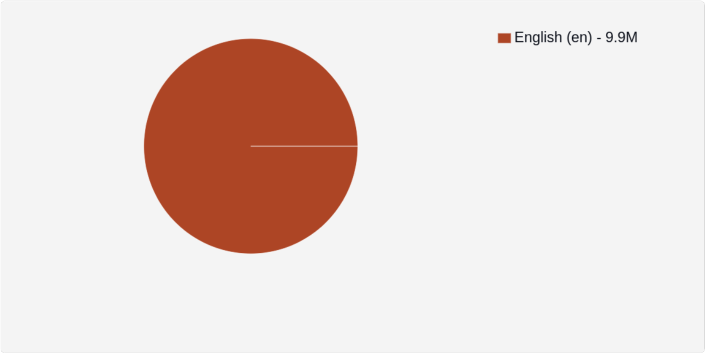
Collections

CC = 71.43%
IA = 28.57%

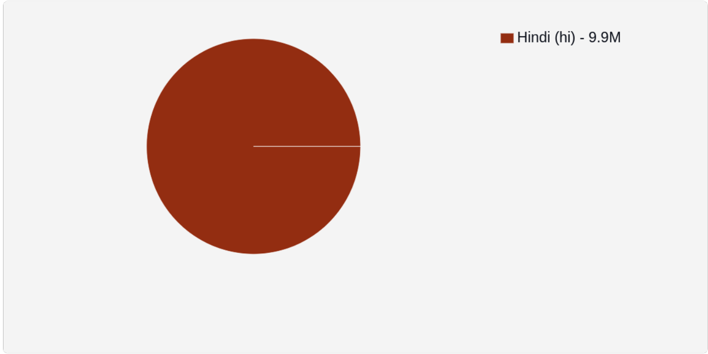


Language Distribution

Source



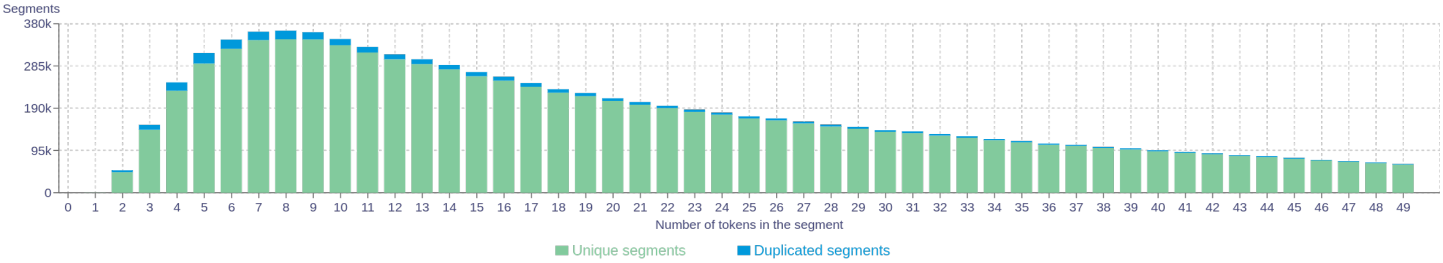
Target



Source segment length distribution by token

<= 49 tokens = **8.4M** segments | **328K** duplicates

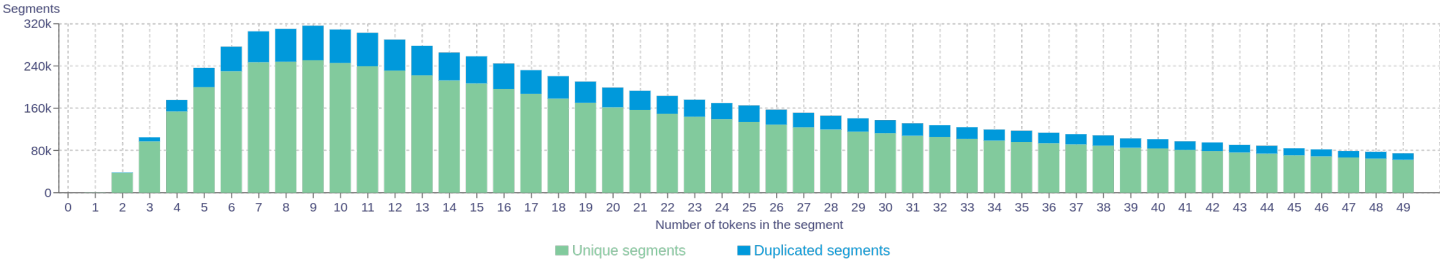
> 50 tokens = **1.1M** segments | **21K** duplicates



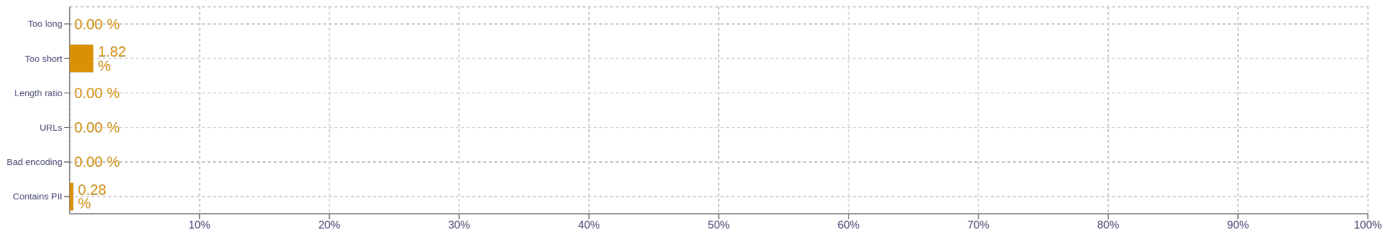
Target segment length distribution by token

<= 49 tokens = **6.6M** segments | **1.5M** duplicates

> 50 tokens = **1.8M** segments | **270K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 490662 one | 400286 india | 373545 new | 369723 use | 365253
2	official website | 71003 admit card | 67980 prime minister | 46063 new delhi | 42350 high quality | 36650
3	visit the official | 20533 bank of india | 17787 jammu and kashmir | 16999 minister narendra modi | 15010 prime minister narendra | 14679
4	visit the official website | 17369 prime minister narendra modi | 14558 wi-fi in public areas | 13346 wi-fi in all rooms | 12087 one of the best | 10005
5	free wi-fi in all rooms | 12084 go to the official website | 8212 streamed directly from their servers | 6709 player fm and our community | 6709 central board of secondary education | 5690

Target n-grams

Size	n-grams
1	भी | 1287608 रूप | 813020 हम | 690171 आपको | 594095 उपयोग | 589754
2	किए गए | 91708 दिए गए | 74571 होना चाहिए | 63561 दी गई | 56566 एडमिट कार्ड | 55688
3	कम से कम | 43579 नीचे दिए गए | 31018 लिंक पर क्लिक | 30258 बटन पर क्लिक | 28712 बारे में अधिक | 26595
4	सार्वजनिक क्षेत्रों में वाई | 15913 बारे में अधिक जानकारी | 14695 रूप में जाना जाता | 11540 स्नानघर में उपलब्ध सुविधाएँ | 10247 भर्ती के लिए आवेदन | 8927
5	काम करने के लिए टेबल | 10187 हमारे समुदाय द्वारा खोजे गए | 6709 सीधे उनके सर्वर से स्ट्रीम | 6709 सर्वर से स्ट्रीम किया जाता | 6709 यूआरएल को अन्य डिजिटल ऑडियो | 6709

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