HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
720,157	19M	100,021,435	95.78 MB

TL tokens	TL characters	TL size
17M	112,959,684	284.92 MB

Dataset top 10 domains

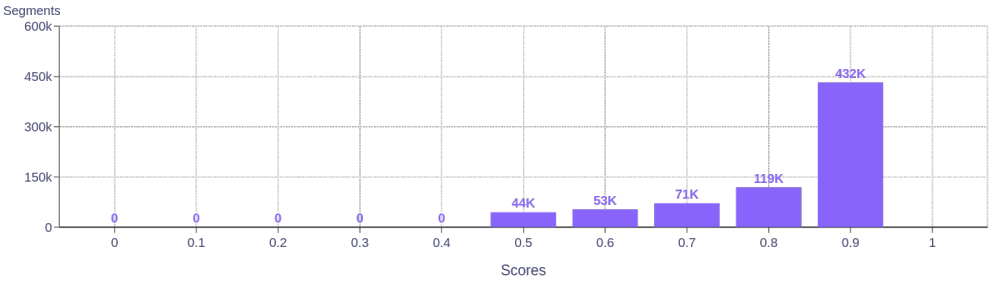
SL domain	Segments	TL domain	Segments
wikipedia.org	31.3%	wikipedia.org	22.3%
hebbarskitchen.com	4.9%	hebbarskitchen.com	4.5%
bajajfinserv.in	4.2%	bajajfinserv.in	3.7%
educationbro.com	3.2%	vsaduidoma.com	2.5%
itsmygame.org	2.9%	itsmygame.org	2.4%
vsaduidoma.com	2.6%	educationbro.com	1.6%
vessoft.com	1.5%	wordproject.org	1.6%
schools-wikipedia.org	1.4%	oneindia.com	1.4%
boldsky.com	1.3%	news18.com	1.4%
uber.com	1.2%	boldsky.com	1.3%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	81.8%	com	64.4%
org	44.9%	org	33.6%
in	10.1%	in	10.1%
net	4.6%	net	4.8%
gov.in	0.8%	gov.in	0.8%
cc	0.6%	nic.in	0.5%
nic.in	0.5%	de	0.4%
co.in	0.5%	co.in	0.4%
plus	0.5%	zone	0.4%
zone	0.4%	news	0.4%

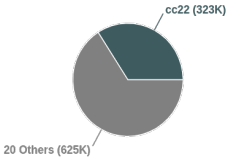
Translation likelihood

≥ 5 = 720K segments | **100.0%**
≥ 8 = 551K segments | **76.6%**
< 5 = 0 segments | **0.0%**



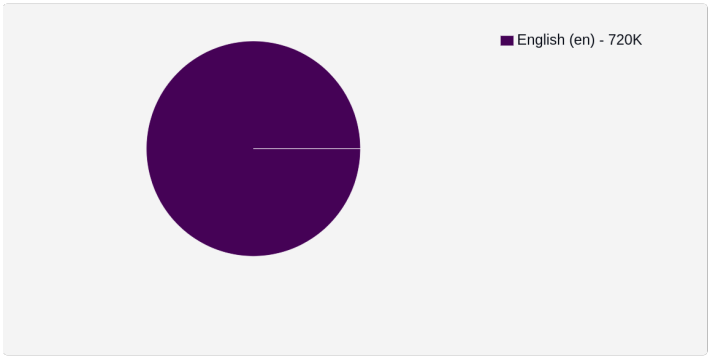
Collections

CC = 66.39%
IA = 33.61%

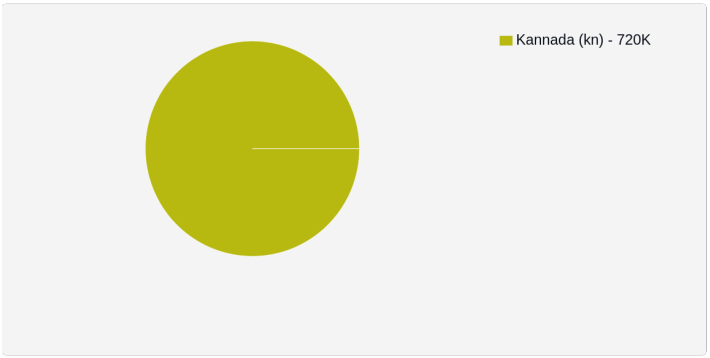


Language Distribution

Source

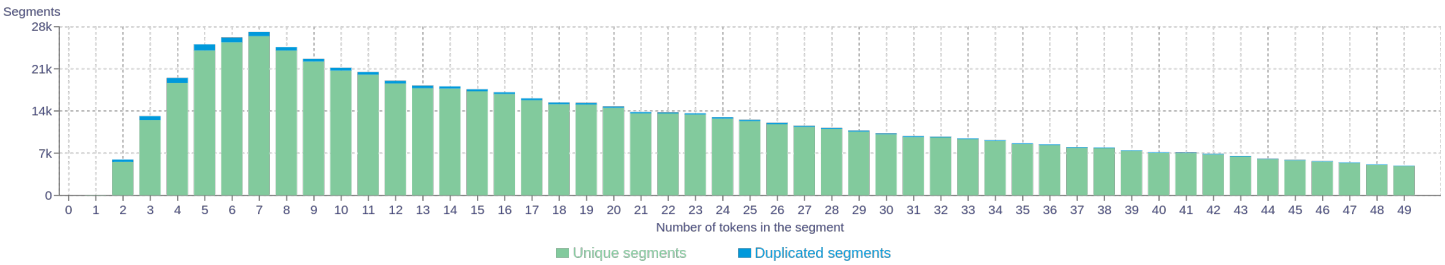


Target



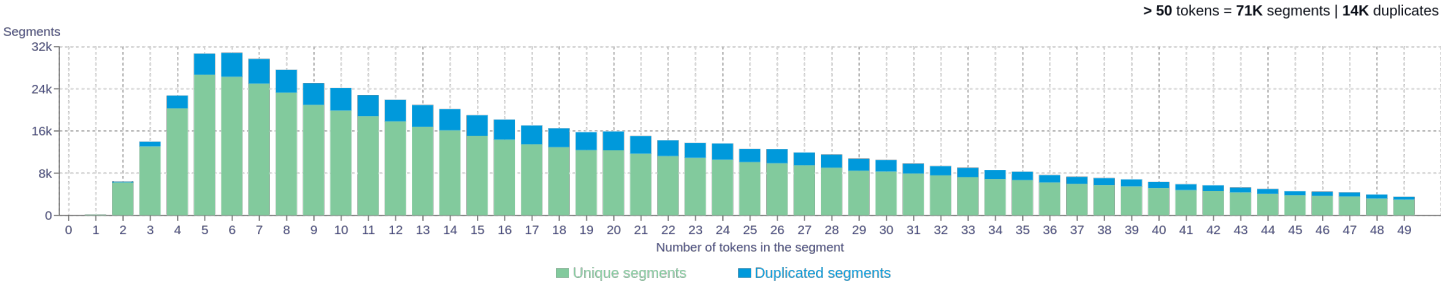
Source segment length distribution by token

<= 49 tokens = **607K** segments | **13K** duplicates
> 50 tokens = **101K** segments | **1.2K** duplicates

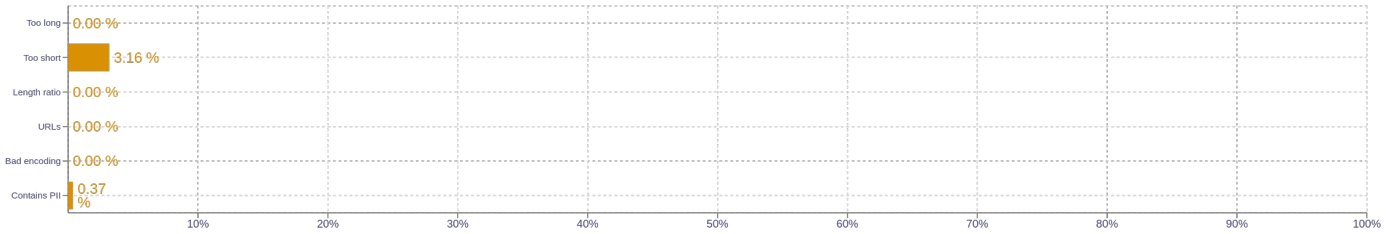


Target segment length distribution by token

<= 49 tokens = **532K** segments | **117K** duplicates
> 50 tokens = **71K** segments | **14K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 37380 · new | 31072 · india | 29940 · one | 29857 · use | 23932
2	new delhi | 5704 · prime minister | 5495 · personal loan | 5004 · bajaj finserv | 4340 · united states | 4113
3	jammu and kashmir | 2270 · minister narendra modi | 2068 · prime minister narendra | 2018 · play the game | 1859 · embed the game | 1621
4	characteristics of the game | 2205 · prime minister narendra modi | 2005 · games like the game | 1026 · step by step photo | 959 · reserve bank of india | 823
5	technical characteristics of the game | 2205 · coupon code to your clipboard | 680 · button to save the coupon | 680 · board of control for cricket | 652 · control for cricket in india | 646

Target n-grams

Size	n-grams
1	ನೀವು | 96932 · ನಿಮ್ಮ | 87887 · ನಾವು | 44718 · ನಮ್ಮ | 36413 · ನಾನು | 28695
2	ಸಹಾಯ ಮಾಡುತ್ತದೆ | 4365 · ಬಜಾಜ್ ಫಿನ್ಸರ್ವ್ | 3091 · ವಿರಾಟ್ ಕೊಹ್ಲಿ | 3071 · ನರೇಂದ್ರ ಮೋದಿ | 3062 · ನಿಮ್ಮ ವೈಯಕ್ತಿಕ | 2964
3	ಪ್ರಧಾನಿ ನರೇಂದ್ರ ಮೋದಿ | 1866 · ನಿಮ್ಮ ವೆಬ್ಪುಟದಲ್ಲಿ ಆಟದ | 1630 · ನಿಮ್ಮ ವೈಯಕ್ತಿಕ ಮಾಹಿತಿಯನ್ನು | 1152 · ಪೆಟ್ರೋಲ್ ಮತ್ತು ಡೀಸೆಲ್ | 1127 · ಜೆನ್ನೈ ಸೂಪರ್ ಕಿಂಗ್ಸ್ | 1118
4	ಕ್ಲಿಪ್ಬೋರ್ಡ್ಗೆ ಕೂಪನ್ ಕೋಡ್ ಉಳಿಸಲು | 751 · ಕೋಡ್ ಉಳಿಸಲು ಗುಂಡಿಯನ್ನು ಕ್ಲಿಕ್ | 751 · ಕೂಪನ್ ಕೋಡ್ ಉಳಿಸಲು ಗುಂಡಿಯನ್ನು | 751 · ಪಾಕವಿಧಾನದ ಈ ಪೋಷ್ಟ್ನೊಂದಿಗೆ ನನ್ನ | 371 · ಬಜಾಜ್ ಫಿನ್ಸರ್ವ್ ಪರ್ಸನಲ್ ಲೋನ್ | 370
5	ಕ್ಲಿಪ್ಬೋರ್ಡ್ಗೆ ಕೂಪನ್ ಕೋಡ್ ಉಳಿಸಲು ಗುಂಡಿಯನ್ನು | 751 · ಕೂಪನ್ ಕೋಡ್ ಉಳಿಸಲು ಗುಂಡಿಯನ್ನು ಕ್ಲಿಕ್ | 751 · ವೈಶಿಷ್ಟ್ಯಗಳನ್ನು ನೋಡಲು ವಿವರಗಳು ಪುಟಕ್ಕೆ ಹೋಗಿ | 246 · ಉತ್ಪನ್ನದ ಕನೆಕ್ಟಿವಿಟಿ ಆಯ್ಕೆಗಳಲ್ಲಿ ಇವುಗಳು ಸೇರಿವೆ | 211 · comನಿಮ್ಮ ಕ್ಲಿಪ್ಬೋರ್ಡ್ಗೆ ಕೂಪನ್ ಕೋಡ್ ಉಳಿಸಲು | 199

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