

General overview

Corpus	Date	Language
hplt-v3-sin_Sinh	9/18/2025	Sinhala (si)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,802,827	38,954,645	27,112,755 (69.60 %)	1.1B	5,938,701,355	14.09 GB

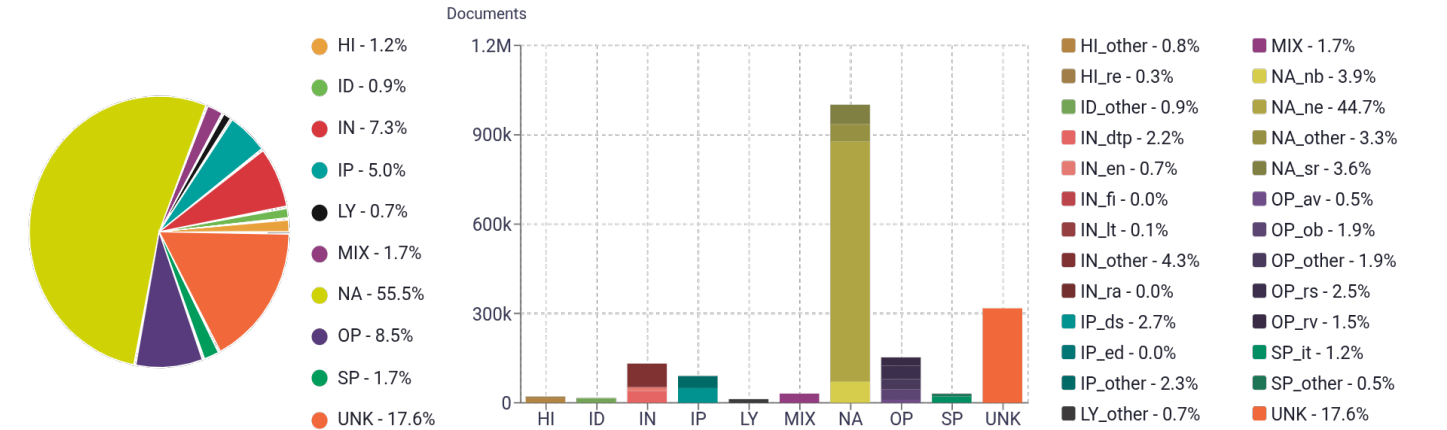
Top 10 domains

Domain	Docs	% of total
blogspot.com	149K	8.29%
lankacnews.com	43K	2.40%
lankadeepa.lk	35K	1.93%
wordpress.com	30K	1.66%
baiscope.lk.com	26K	1.46%
divaina.lk	20K	1.08%
itnnews.lk	17K	0.93%
thepapare.com	15K	0.82%
mawbima.lk	14K	0.79%
theleader.lk	14K	0.78%

Top 10 TLDs

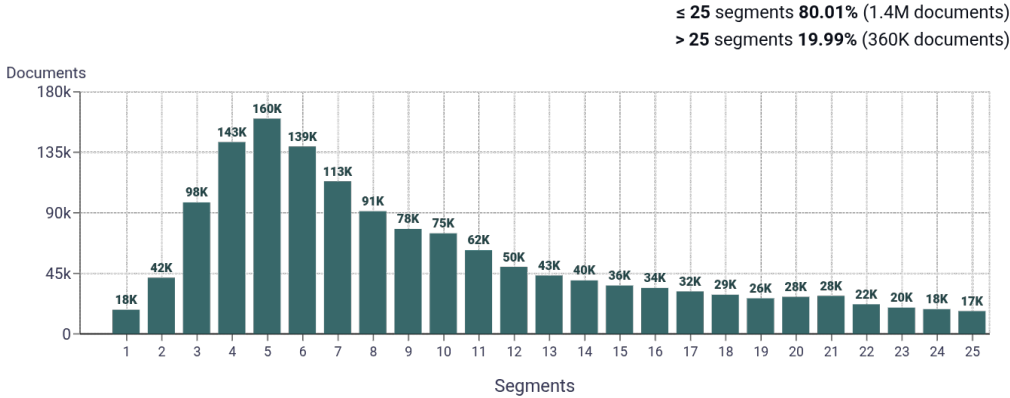
Domain	Docs	% of total
com	861K	47.74%
lk	664K	36.86%
org	75K	4.16%
net	48K	2.64%
info	25K	1.36%
gov.lk	17K	0.94%
media	9.5K	0.52%
us	6.8K	0.38%
co.uk	6.7K	0.37%
cn	5.8K	0.32%

Register labels

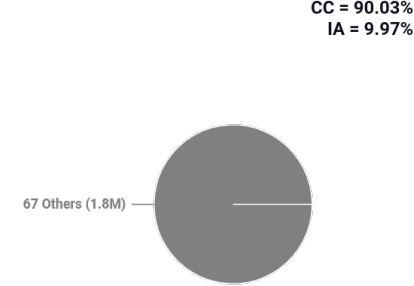


MT:10.6% | 192K Documents

Documents size (in segments) ⓘ

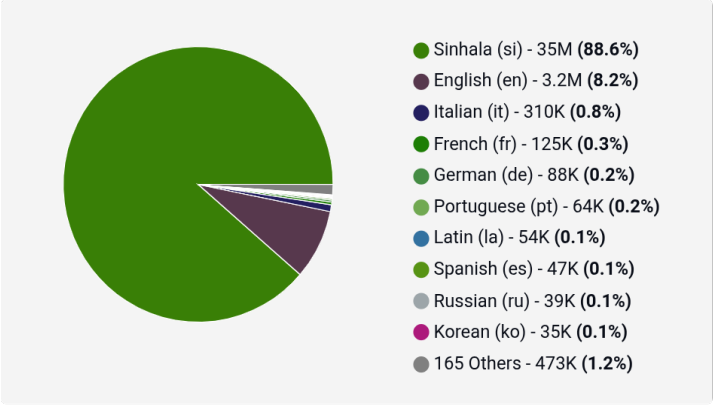


Document collections

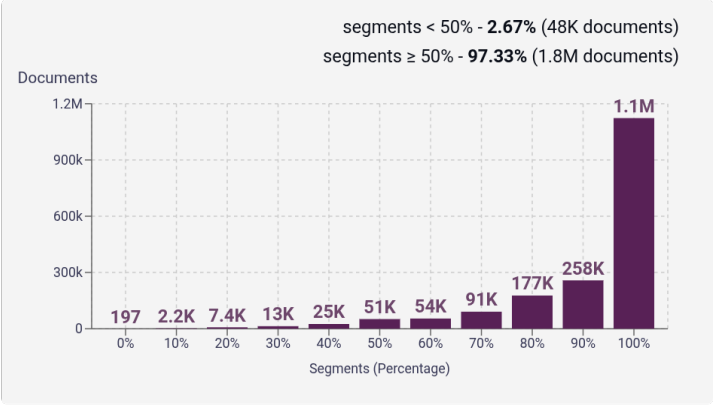


Language Distribution

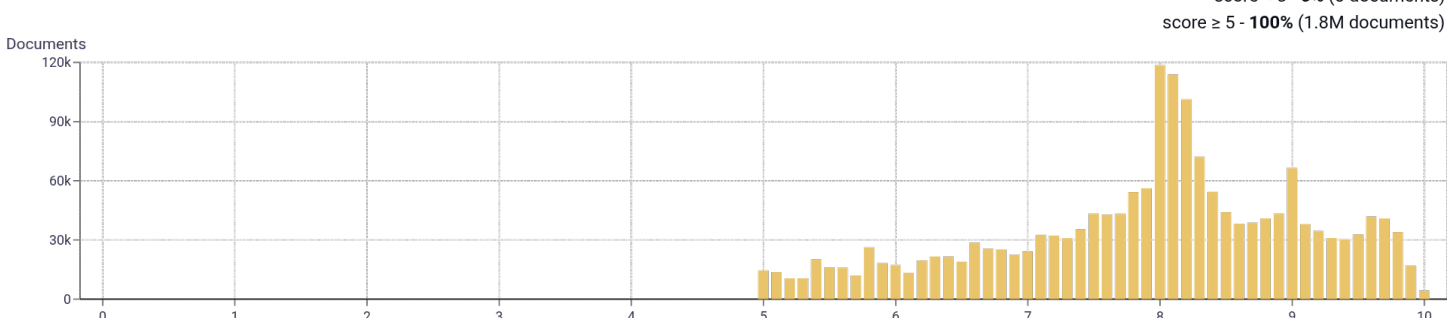
Number of segments in the Sinhala (si) corpus



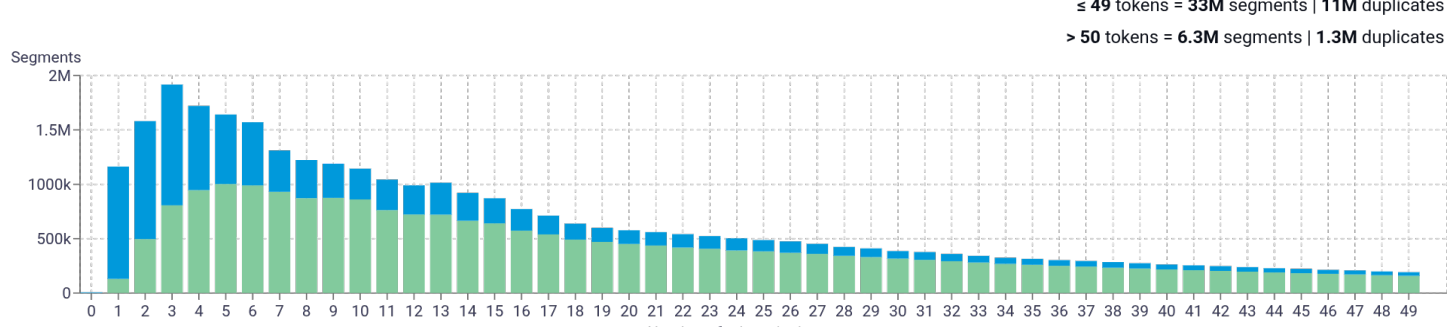
Percentage of segments in Sinhala (si) inside documents



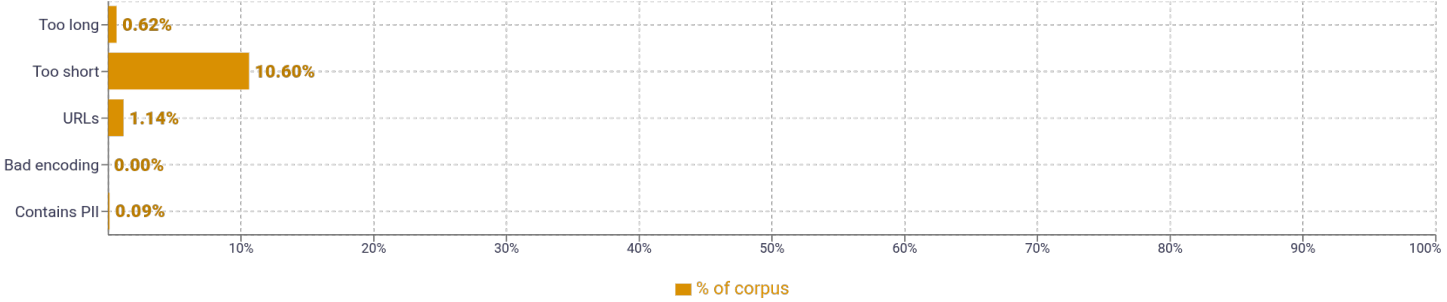
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>කර 3,680,182</div> <div>කරන 3,193,992</div> <div>කළ 2,918,195</div> <div>මම 2,686,107</div> <div>එක 2,679,596</div>	
2	<div>ශ්‍රී ලංකා 706,689</div> <div>කරන ලද 432,864</div> <div>කර ඇත 378,231</div> <div>ශ්‍රී ලංකාවේ 293,336</div> <div>කළ හැකි 288,969</div>	
3	<div>ශ්‍රී ලංකා නිදහස් 67,404</div> <div>ශ්‍රී ලංකා ක්‍රිකට් 62,405</div> <div>post a comment 59,847</div> <div>රනිල් වික්‍රමසිංහ මහතා 55,892</div> <div>ලබා ගත හැකි 52,501</div>	
4	<div>ශ්‍රී ලංකා නිදහස් පක්ෂයේ 22,796</div> <div>ශ්‍රී ලංකා නිදහස් පක්ෂය 21,890</div> <div>ජනාධිපති රනිල් වික්‍රමසිංහ මහතා 20,804</div> <div>due to copyright issues 18,485</div> <div>we do not provide 16,955</div>	
5	<div>we do not provide any 16,912</div> <div>do not provide any torrent 16,779</div> <div>අදහස් පළ කිරීමට ජ්‍යෙෂ්ඨ වන්න 16,302</div> <div>not provide any torrent links 14,236</div> <div>comment has been removed by 13,172</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				