HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-khm_Khmr	9/18/2025	Khmer (km)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,323,657	20,380,572	13,736,261 (67.40 %)	3.4B	4,964,306,032	12.84 GB

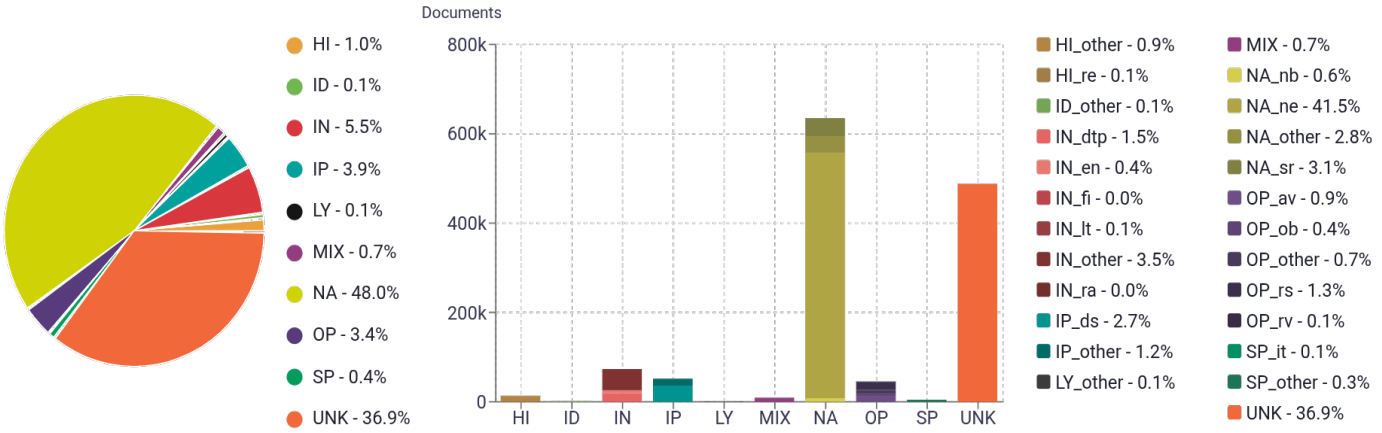
Top 10 domains

Domain	Docs	% of total
kohsantepheapda...	52K	3.91%
freshnewsasia.com	31K	2.31%
rasmeinews.com	26K	1.95%
wordpress.com	22K	1.65%
khtoem.com	21K	1.57%
khmerload.com	19K	1.44%
dap-news.com	18K	1.36%
postnews.com.kh	14K	1.06%
kampucheathmey.com	13K	0.98%
nokorwatnews.com	13K	0.95%

Top 10 TLDs

Domain	Docs	% of total
com	560K	42.28%
icu	368K	27.82%
com.kh	140K	10.56%
gov.kh	51K	3.86%
org	44K	3.29%
net	28K	2.10%
vn	24K	1.85%
news	19K	1.47%
info	18K	1.37%
org.kh	16K	1.23%

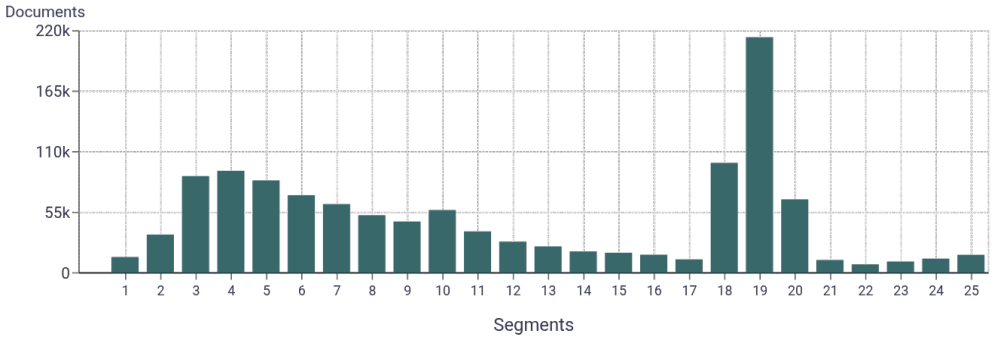
Register labels



MT:33.6% | 444K Documents

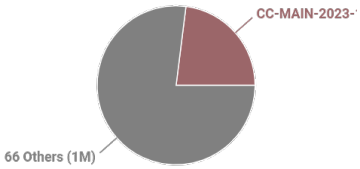
Documents size (in segments)

≤ 25 segments **90.71%** (1.2M documents)
> 25 segments **9.29%** (123K documents)



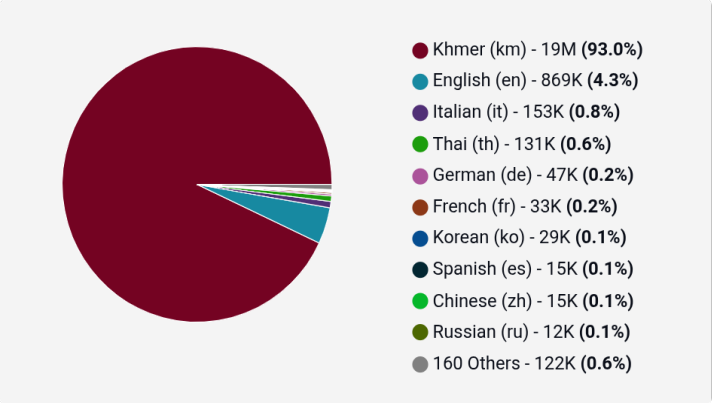
Document collections

CC = 92.44%
IA = 7.56%

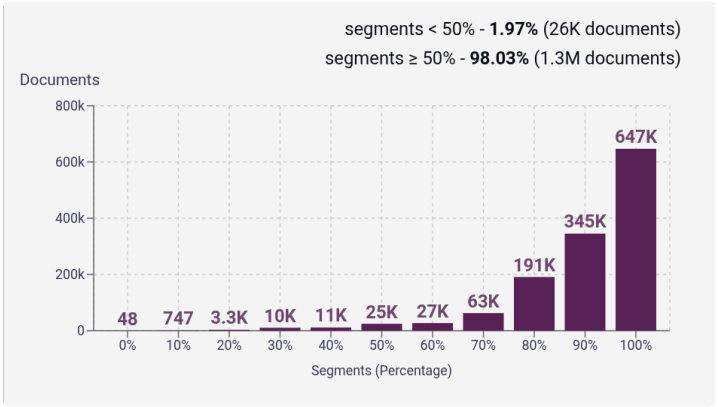


Language Distribution

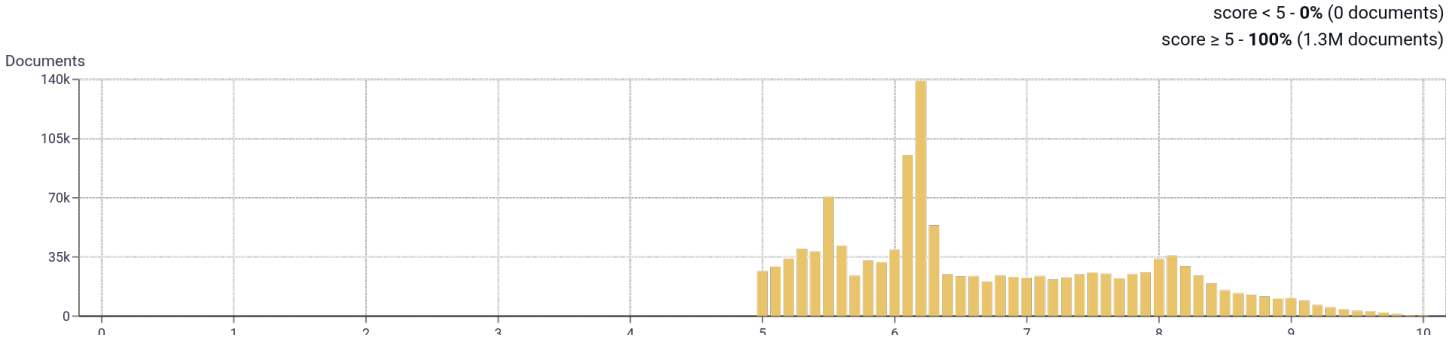
Number of segments in the Khmer (km) corpus



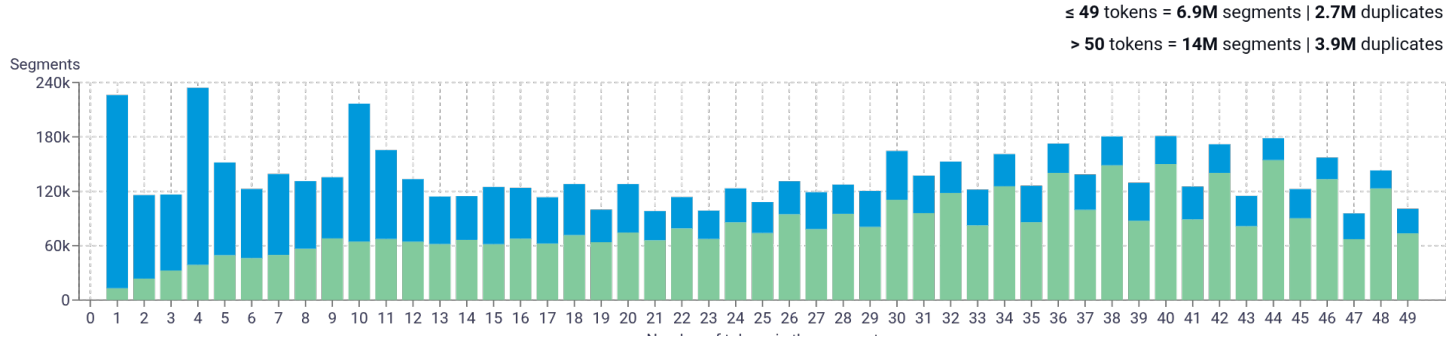
Percentage of segments in Khmer (km) inside documents



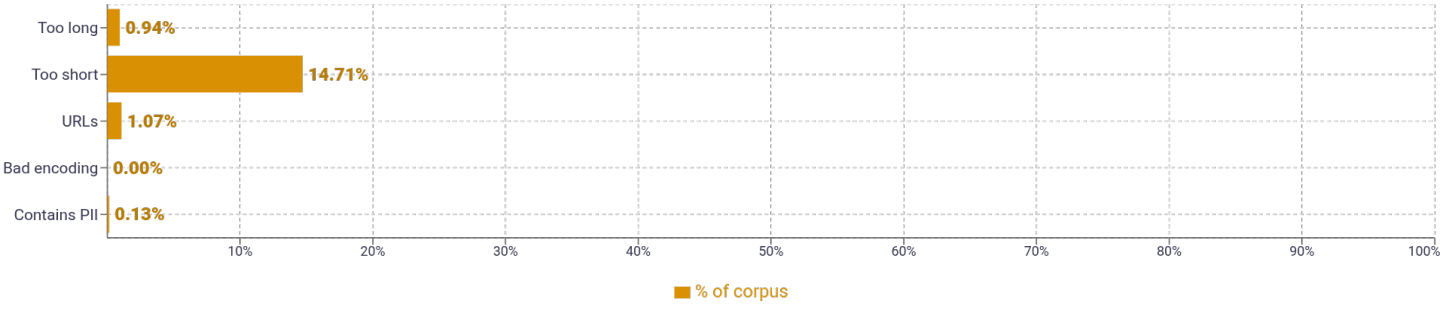
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ល | 107,887,753 · ក | 102,038,688 · ន | 96,713,435 · រ | 91,071,135 · ប | 71,494,724
2	ត ក | 1,031,266 · ន ន | 697,076 · ញ ប | 677,098 · ញ ក | 644,537 · រ ន | 580,949
3	jackpot party slots | 434,681 · sic bo ដ | 253,972 · jackpot slots ឥតគិត | 217,526 · party slots ស | 121,558 · baccarat online ដ | 116,938
4	jackpot party slots ស | 118,867 · ឯ baccarat online ដ | 116,654 · free credit casino no | 111,011 · credit casino no deposit | 111,011 · party casino slots ឥតគិត | 110,893
5	free credit casino no deposit | 111,011 · cambodia free credit casino no | 109,080 · jackpot party casino slots ឥតគិត | 108,966 · slot machine site games ក | 107,370 · ញ mega jackpot party slots | 102,841

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

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				