

General overview

Corpus	Date	Language
hplt-v3-kam_Latn	9/17/2025	Kamba (kam)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,043	13,065	12,103 (92.64 %)	714K	3,619,686	3.86 MB

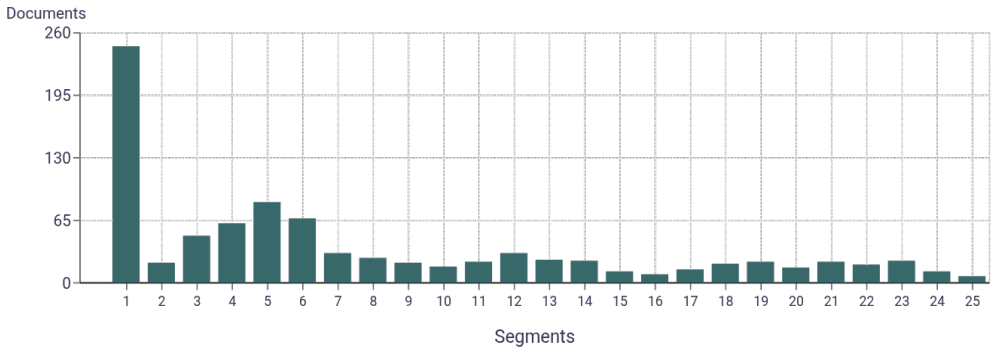
Top 10 domains

Domain	Docs	% of total
jw.org	536	51.39%
bible.is	244	23.39%
rmsradio.co.ke	107	10.26%
sangufm.co.ke	75	7.19%
gotquestions.org	10	0.96%
sharptipnews.com	8	0.77%
gafkosoft.com	8	0.77%
bible.com	8	0.77%
kbc.co.ke	7	0.67%
kituionline.com	6	0.58%

Top 10 TLDs

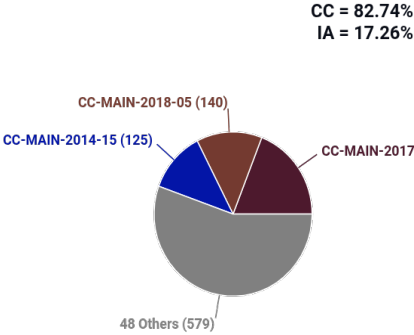
Domain	Docs	% of total
org	562	53.88%
is	244	23.39%
co.ke	192	18.41%
com	42	4.03%
net	2	0.19%
web.id	1	0.10%

Documents size (in segments) ⓘ



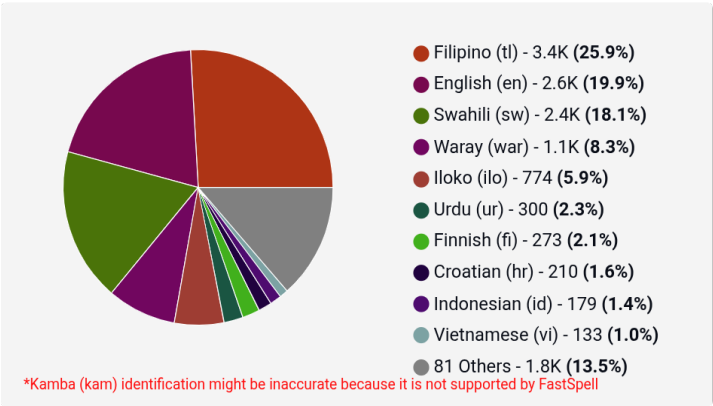
≤ 25 segments **86.29%** (900 documents)  
> 25 segments **13.71%** (143 documents)

Document collections

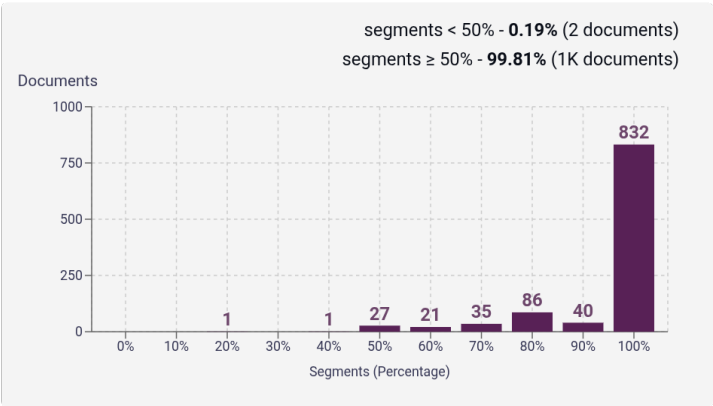


Language Distribution

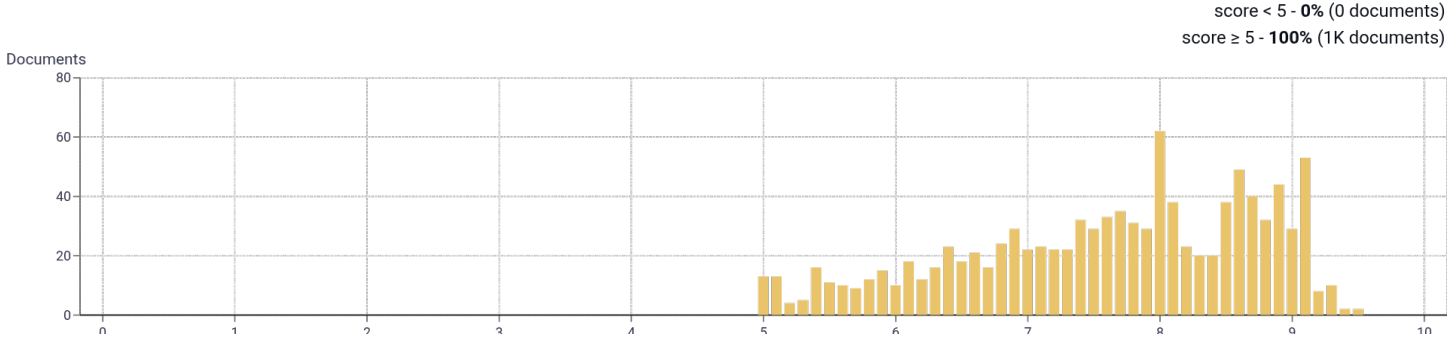
Number of segments in the Kamba (kam) corpus



Percentage of segments in Kamba (kam) inside documents

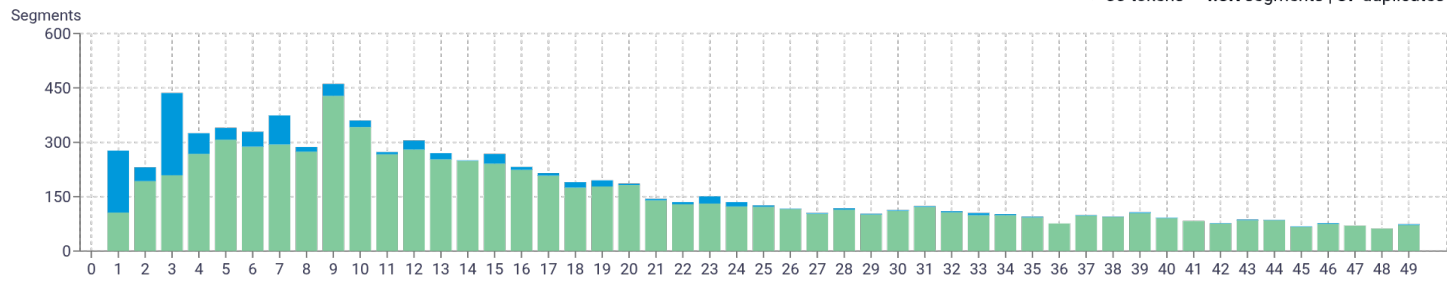


Distribution of documents by document score

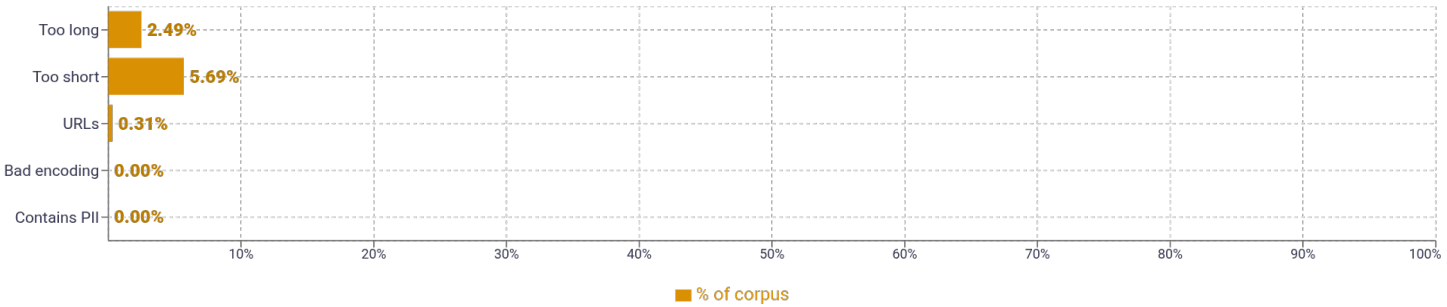


Segment length distribution by token

≤ 49 tokens = 8.8K segments | 926 duplicates  
> 50 tokens = 4.3K segments | 37 duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	andũ   4,083   yeova   4,060   kwa   3,745   ĩndĩ   2,780   ĩla   2,545	📄
2	kwa ngelekany   401   andũ aingĩ   327   kũũ nthĩ   318   kwa nzĩa   318   sya yeova   280	📄
3	ĩũũ wa nthĩ   311   ngũsĩ sya yeova   263   tene na tene   247   meko ma atũmwa   138   atamu na eva   138	📄
4	ũu wĩ o vo   65   ĩndĩ o na ũu   50   ũvoo mũseo wa ũsumbĩ   45   maũndũ ma vata kuma   41   ĩũũ wa nthĩ yonthe   38	📄
5	ĩndĩ o na ũu wĩ   47   maũndũ ma vata kuma ndetonĩ   40   kũsoma mbivĩlia kya kyumwa kĩĩ   34   kya kũsoma mbivĩlia kya kyumwa   34	📄
	the holy bible in current   29	

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				