HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-kik_Latn	9/17/2025	Kikuyu

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
8,625	111,864	88,402 (79.03 %)	3.6M	17,885,135	19.55 MB

Top 10 domains

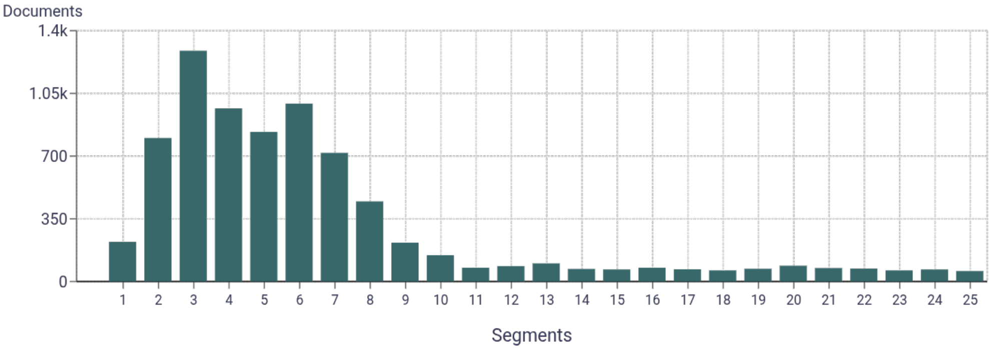
Domain	Docs	% of total
mediamaxnetwork...	2.7K	31.29%
radiokimuri.co.ke	1.5K	17.25%
kayufm.com	986	11.43%
biblica.com	958	11.11%
jw.org	510	5.91%
rmsradio.co.ke	313	3.63%
wikipedia.org	201	2.33%
inoorotv.co.ke	195	2.26%
purekikuyulyric...	137	1.59%
blogspot.com	125	1.45%

Top 10 TLDs

Domain	Docs	% of total
co.ke	5.2K	60.56%
com	2.5K	29.10%
org	824	9.55%
is	35	0.41%
jp	18	0.21%
eu	3	0.03%
report	2	0.02%
net	2	0.02%
tw	1	0.01%
ru	1	0.01%

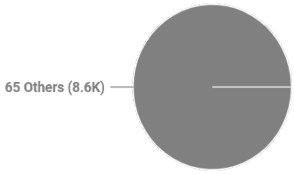
Documents size (in segments)

≤ 25 segments **89.67%** (7.7K documents)
> 25 segments **10.33%** (891 documents)



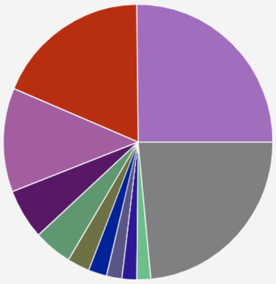
Document collections

CC = 92.56%
IA = 7.44%



Language Distribution

Number of segments in the Kikuyu corpus

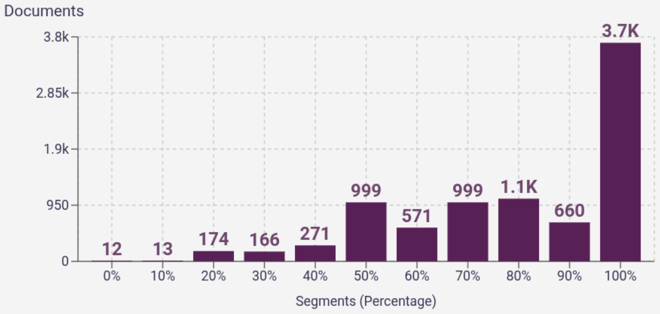


- ● English - 28K **(25.2%)**
- ● Swahili - 21K **(18.5%)**
- ● Filipino - 14K **(12.2%)**
- ● Amharic - 6.6K **(5.9%)**
- ● Waray - 5.1K **(4.6%)**
- ● Iloko - 3K **(2.7%)**
- ● Korean - 2.5K **(2.2%)**
- ● Croatian - 2.1K **(1.8%)**
- ● Spanish - 2K **(1.8%)**
- ● Esperanto - 1.9K **(1.7%)**
- ● 135 Others - 26K **(23.5%)**

*Kikuyu identification might be inaccurate because it is not supported by FastSpell.

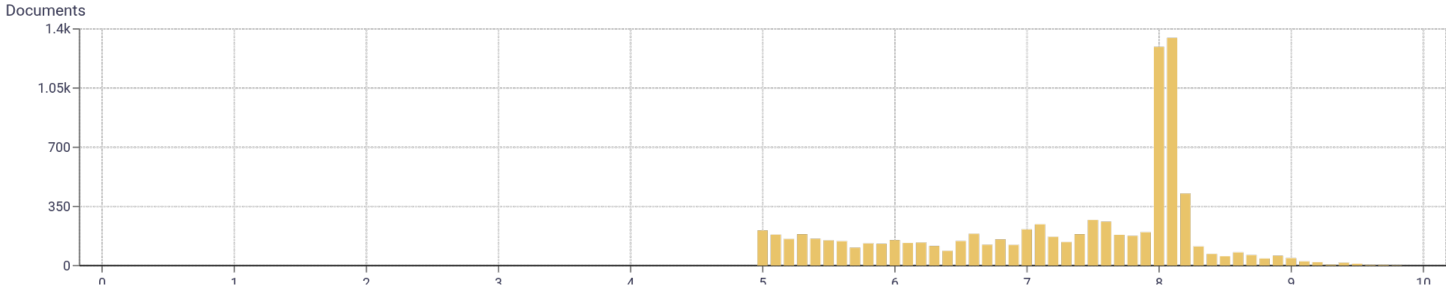
Percentage of segments in Kikuyu inside documents

segments < 50% - **7.37%** (636 documents)
segments ≥ 50% - **92.63%** (8K documents)



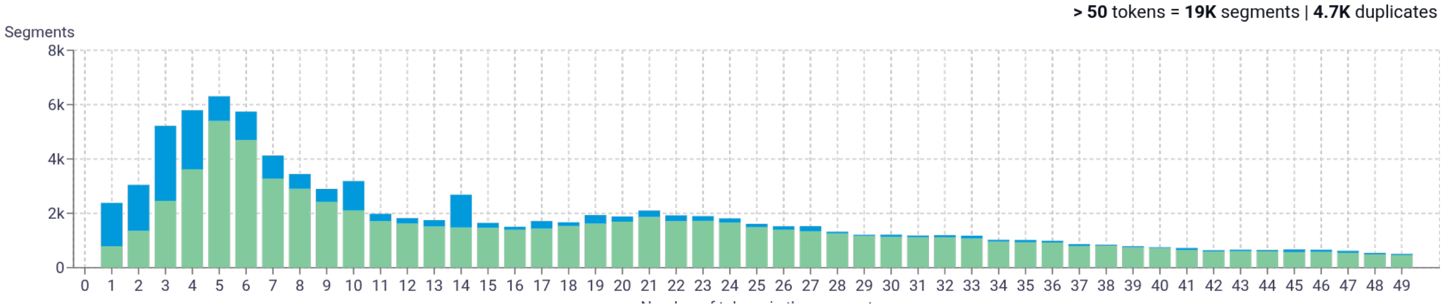
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (8.6K documents)

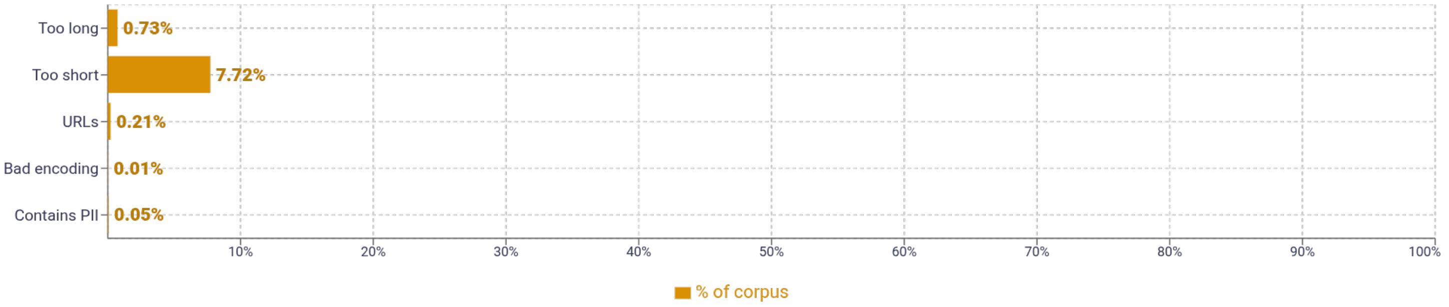


Segment length distribution by token

≤ **49** tokens = **93K** segments | **19K** duplicates
> 50 tokens = **19K** segments | **4.7K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	wa | 91,666 cia | 18,562 ini | 15,612 ngai | 13,632 inĩ | 12,791
2	ũndũ wa | 3,980 thutha wa | 3,496 thĩinĩ wa | 3,429 mwena wa | 3,103 ũndũ ũcio | 3,035
3	mt kenya region | 985 mbica irutitwe online | 955 mutongoria wa bururi | 913 written by kayu | 844 kayu digital team | 836
4	written by kayu digital | 836 kayu digital team on | 836 by kayu digital team | 836 mutongoria wa bururi uhuru | 489 wa bururi uhuru kenyatta | 468
5	written by kayu digital team | 836 by kayu digital team on | 836 mutongoria wa bururi uhuru kenyatta | 463 ariũ na aarĩ a ithe | 288 muoyo wa tene na tene | 221

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				