# HPLT Analytics report

**HPLT**Analytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| kik_Latn.jsonl.tsv | 11/5/2024 | Kikuyu (ki) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 3,995 | 51,931 | 33,013 (63.57 %) | 1.8M | 9.92 MB | 9,244,068 |

### Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| mediamaxnetwork.co.ke | 1.6K | 40.48 |
| jw.org | 713 | 17.85 |
| wikipedia.org | 355 | 8.89 |
| bible.is | 224 | 5.61 |
| biblica.com | 191 | 4.78 |
| blogspot.jp | 91 | 2.28 |
| thekenyatoday.com | 69 | 1.73 |
| blogspot.com | 66 | 1.65 |
| radiokimuri.co.ke | 64 | 1.60 |
| blogspot.fr | 38 | 0.95 |

### Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| co.ke | 1.9K | 47.53 |
| org | 1.2K | 28.96 |
| com | 540 | 13.52 |
| is | 224 | 5.61 |
| jp | 91 | 2.28 |
| fr | 43 | 1.08 |
| de | 15 | 0.38 |
| net | 13 | 0.33 |
| eu | 3 | 0.08 |
| click | 1 | 0.03 |

## Documents size (in segments)

<= 25 segments **87.83%** (3.5K documents)
> 25 segments **12.17%** (485 documents)
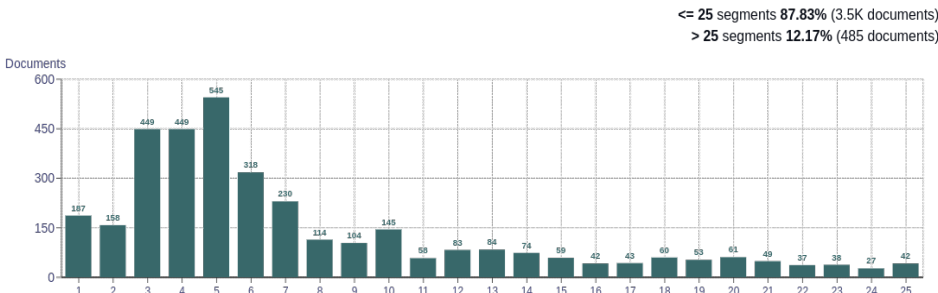


## Documents by collection

cc18 (1.7K)
wide16 (465)
19 Others (1.8K)
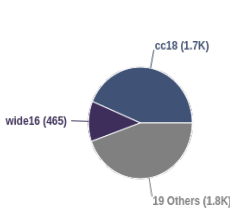


## Language Distribution

### Number of segments

- English (en) - 12K
- Filipino (tl) - 8.7K
- Swahili (sw) - 7.9K
- Waray (war) - 3.1K
- Iloko (ilo) - 2.3K
- Vietnamese (vi) - 1.4K
- Italian (it) - 1.2K
- Croatian (hr) - 1.1K
- Spanish (es) - 939
- Latin (la) - 851
- 119 Others - 12K

*Kikuyu (ki) identification might be inaccurate because language is not supported by Fasttext
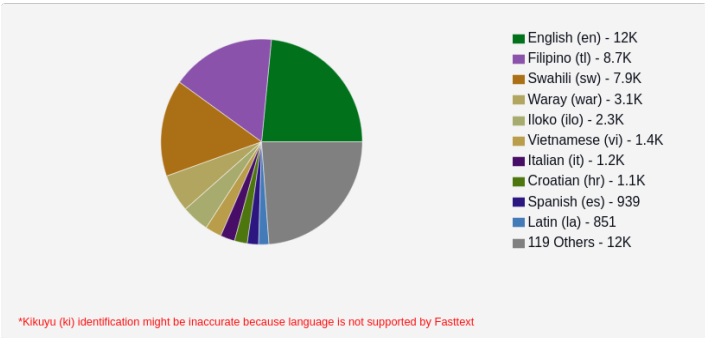


### Percentage of segments in Kikuyu (ki) inside documents
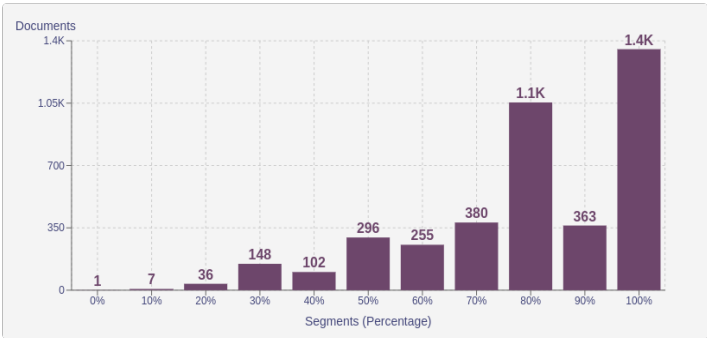


| Segments (Percentage) | Documents |
|---|---|
| 0% | 1 |
| 10% | 7 |
| 20% | 36 |
| 30% | 148 |
| 40% | 102 |
| 50% | 296 |
| 60% | 255 |
| 70% | 380 |
| 80% | 1.1K |
| 90% | 363 |
| 100% | 1.4K |

## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (4K documents)



## Segment length distribution by token

<= 49 tokens = **26K** segments | **16K** duplicates
> 50 tokens = **10K** segments | **2.8K** duplicates



- Unique segments
- Duplicated segments

## Segment noise distribution

| | |
|---|---|
| Too long | 0.00 % |
| Too short | 9.34 % |
| URLs | 0.42 % |
| Bad encoding | 0.00 % |
| Contains PII | 0.02 % |
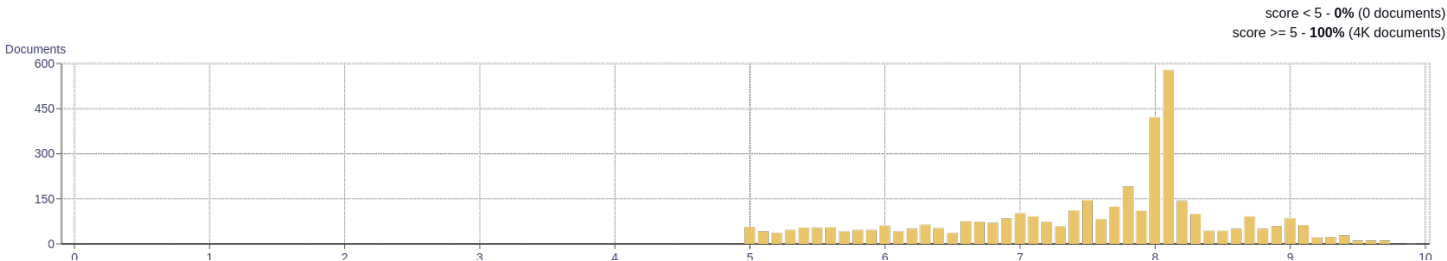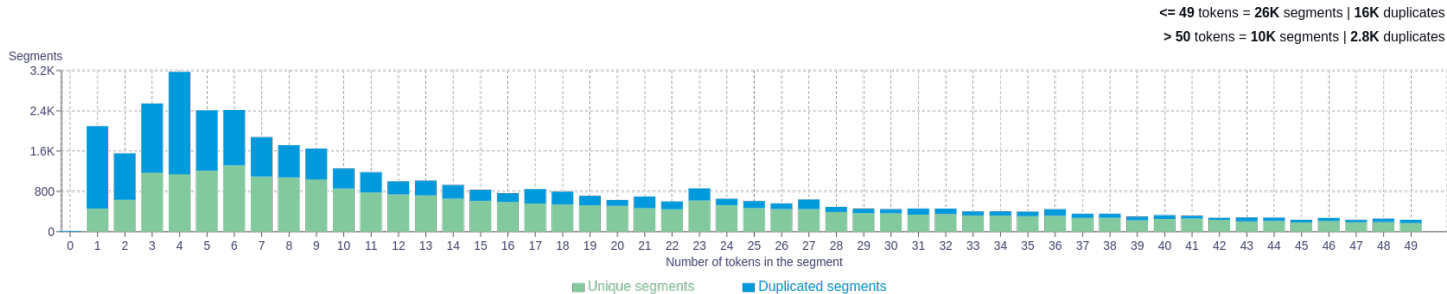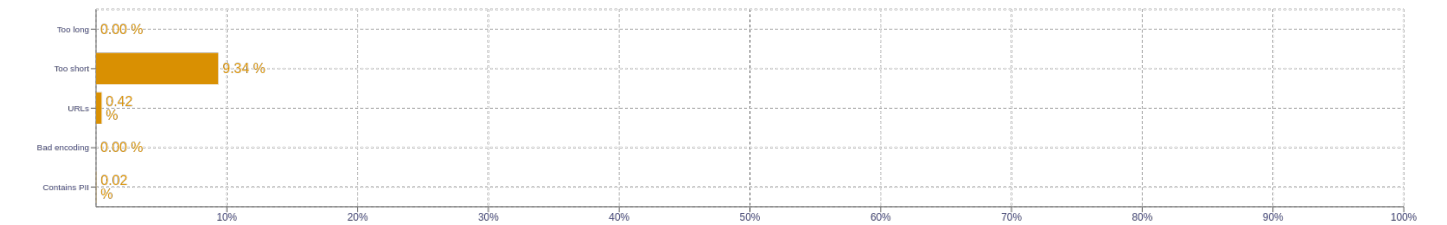
## About HPLT Analytics

**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