# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| sag_Latn.jsonl.tsv | 11/27/2024 | Sango (sg) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 3,161 | 51,900 | 31,786 (61.24 %) | 4.2M | 16.56 MB | 16,692,086 |

## Top 10 domains

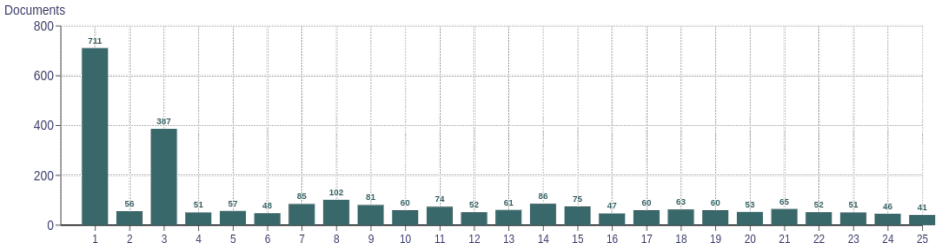
| Domain | Docs | % of total |
|---|---|---|
| jw.org | 2K | 63.08 |
| bible.is | 796 | 25.18 |
| wikipedia.org | 66 | 2.09 |
| islamhouse.com | 42 | 1.33 |
| gotquestions.org | 26 | 0.82 |
| icc-cpi.int | 15 | 0.47 |
| siriri.org | 15 | 0.47 |
| free.fr | 11 | 0.35 |
| wiktionary.org | 11 | 0.35 |
| ebible.org | 11 | 0.35 |

## Top 10 TLDs

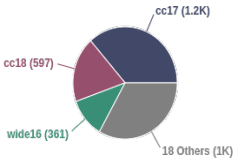
| Domain | Docs | % of total |
|---|---|---|
| org | 2.2K | 68.14 |
| is | 796 | 25.18 |
| com | 132 | 4.18 |
| fr | 31 | 0.98 |
| int | 15 | 0.47 |
| net | 9 | 0.28 |
| blog | 4 | 0.13 |
| info | 4 | 0.13 |
| ch | 3 | 0.09 |
| eu | 2 | 0.06 |

## Documents size (in segments)

**<= 25** segments **79.85%** (2.5K documents)
**> 25** segments **20.15%** (637 documents)

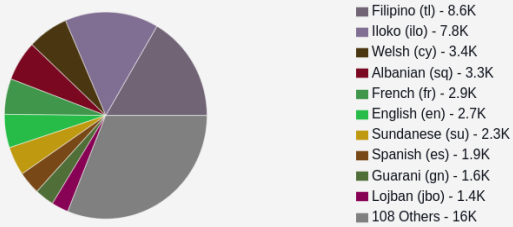



## Documents by collection



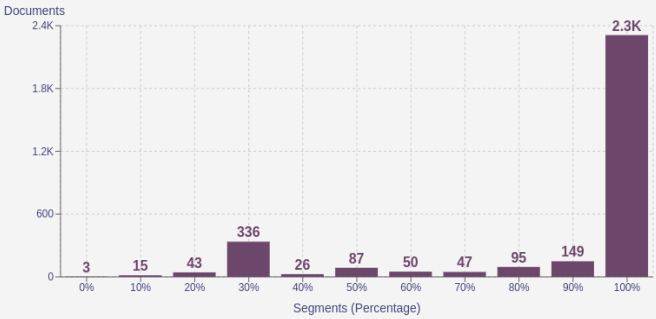

## Language Distribution

### Number of segments



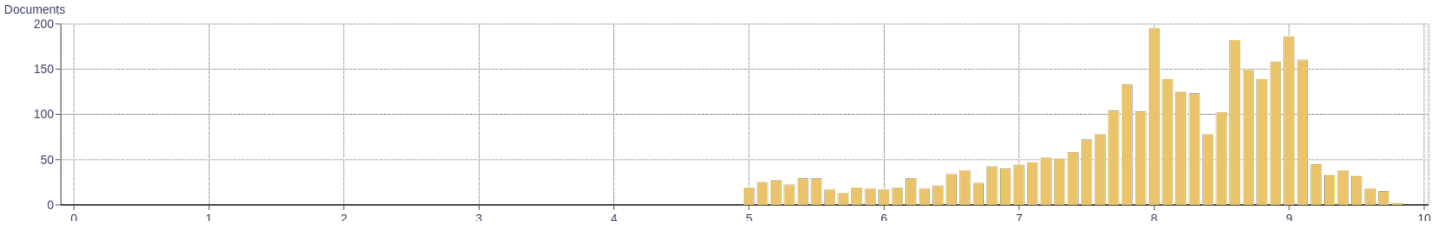

*Sango (sg) identification might be inaccurate because language is not supported by Fasttext

### Percentage of segments in Sango (sg) inside documents



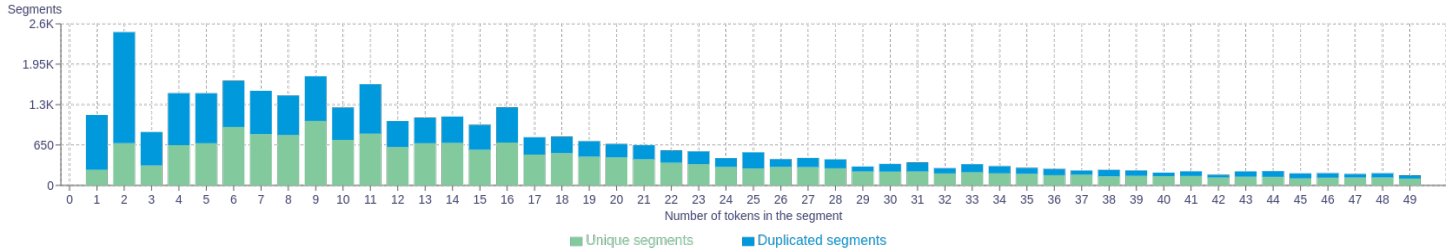

## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.2K documents)

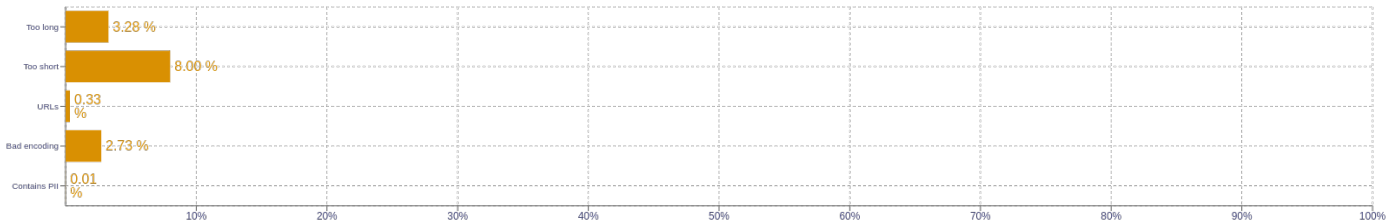



## Segment length distribution by token

**<= 49** tokens = **19K** segments | **15K** duplicates
**> 50** tokens = **18K** segments | **4.6K** duplicates




## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | ala \| 77892 · ayeke \| 67232 · yeke \| 51236 · ye \| 50227 · mbi \| 35944 |
| 2 | ala yeke \| 8886 · tongana nyen \| 6043 · mbi yeke \| 5683 · yeke sara \| 4840 · sara ye \| 4306 |
| 3 | ngbanga ti nyen \| 3107 · ye so ayeke \| 1931 · azo so ayeke \| 1851 · ayeke na yâ \| 1811 · tere ti ala \| 1643 |
| 4 | kobe kobe kobe kobe \| 963 · 9article 9article 9article 9article \| 716 · tour ti ba ndo \| 598 · alingbi ti mû maboko \| 575 · ape ape ape ape \| 569 |
| 5 | kobe kobe kobe kobe kobe \| 962 · 9article 9article 9article 9article 9article \| 712 · ape ape ape ape ape \| 566 · nâr bi afvike yâ rahmân \| 494 · ayeke na yâ ti bible \| 374 |

## 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