HPLT Analytics report

General overview

Corpus	Analytics date	Language
luo_Latn.jsonl.tsv	12/6/2024	Luo (luo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
4,153	84,116	52,703 (62.66 %)	4.6M	19.67 MB	20,245,474

Top 10 domains

Domain	Docs	% of total
bible.is	1.3K	31.28
jw.org	1.1K	26.01
bibles.org	970	23.36
jaluo.com	116	2.79
blogspot.com	59	1.42
skyfm.co.ke	53	1.28
blogspot.co.ke	40	0.96
egwwritings.org	39	0.94
ebible.org	31	0.75
dwonlwak.com	31	0.75

Top 10 TLDs

Domain	Docs	% of total
org	2.2K	54.06
is	1.3K	31.28
com	390	9.39
co.ke	115	2.77
net	28	0.67
bible	15	0.36
com.br	11	0.26
fr	8	0.19
co.uk	5	0.12
info	5	0.12

Documents size (in segments)

<= 25 segments **83.27%** (3.5K documents)
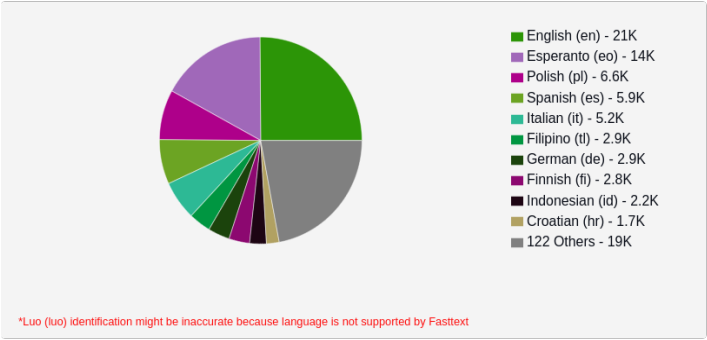> 25 segments **16.73%** (695 documents)
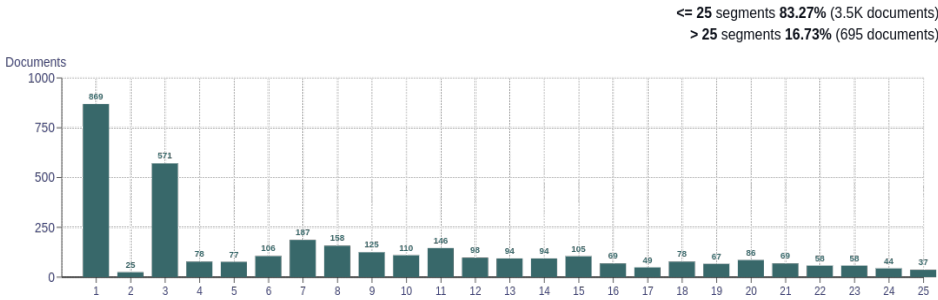
Documents by collection
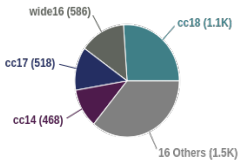
Language Distribution

Number of segments

- English (en) - 21K
- Esperanto (eo) - 14K
- Polish (pl) - 6.6K
- Spanish (es) - 5.9K
- Italian (it) - 5.2K
- Filipino (tl) - 2.9K
- German (de) - 2.9K
- Finnish (fi) - 2.8K
- Indonesian (id) - 2.2K
- Croatian (hr) - 1.7K
- 122 Others - 19K

*Luo (luo) identification might be inaccurate because language is not supported by Fasttext
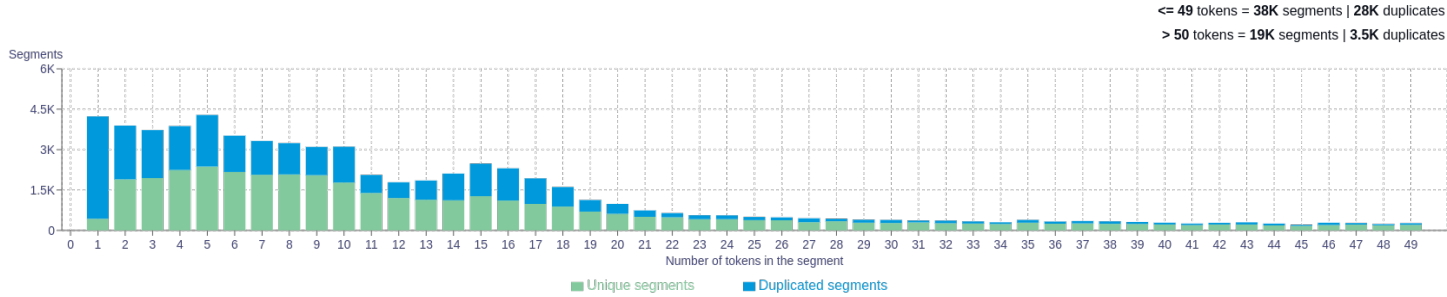
Percentage of segments in Luo (luo) inside documents
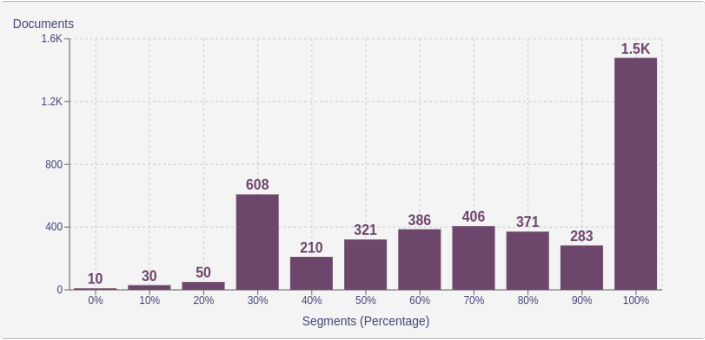
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (4.2K documents)

Segment length distribution by token

<= 49 tokens = **38K** segments | **28K** duplicates
> 50 tokens = **19K** segments | **3.5K** duplicates

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	3.39 %
Too short	12.75 %
URLs	0.66 %
Bad encoding	0.02 %
Contains PII	0.22 %
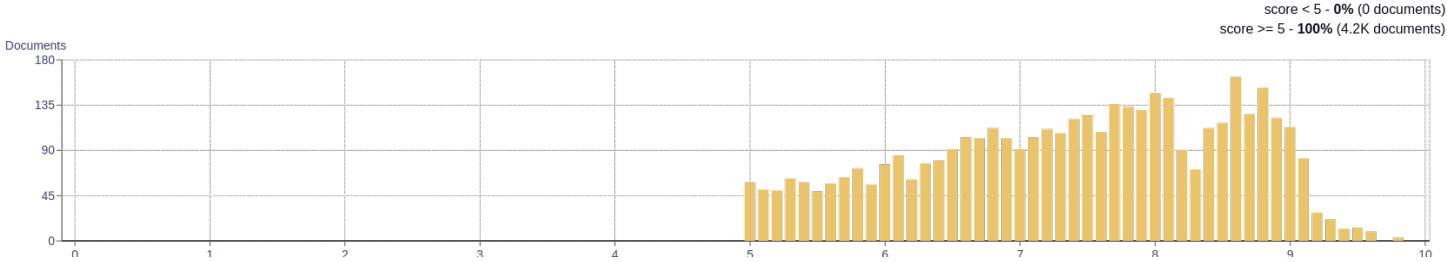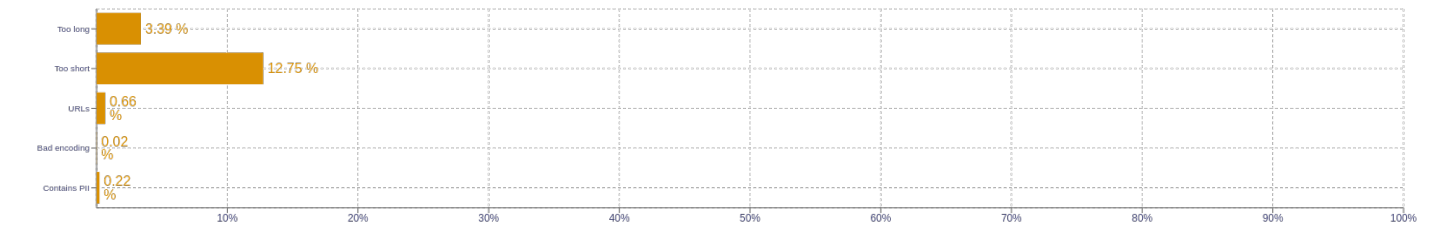
Frequent n-grams

Size	n-grams
1	morite | 181169 i | 84822 ni | 70848 ki | 53027 e | 46982
2	morite morite | 180621 ok ok | 12787 i kom | 8591 ki i | 7545 nono nono | 4364
3	morite morite morite | 180075 ok ok ok | 12529 nono nono nono | 4184 woko ki i | 1520 i ot pa | 908
4	morite morite morite morite | 179529 ok ok ok ok | 12327 nono nono nono nono | 4091 moritemorite morite morite morite | 540 morite moritemorite morite morite | 540
5	morite morite morite morite morite | 178983 ok ok ok ok ok | 12148 nono nono nono nono nono | 4052 moritemorite morite morite morite morite | 540 morite moritemorite morite morite morite | 540

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