HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-luo_Latn	9/17/2025	Luo

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,611	103,973	92,927 (89.38 %)	4.8M	21,112,036	20.48 MB

Top 10 domains

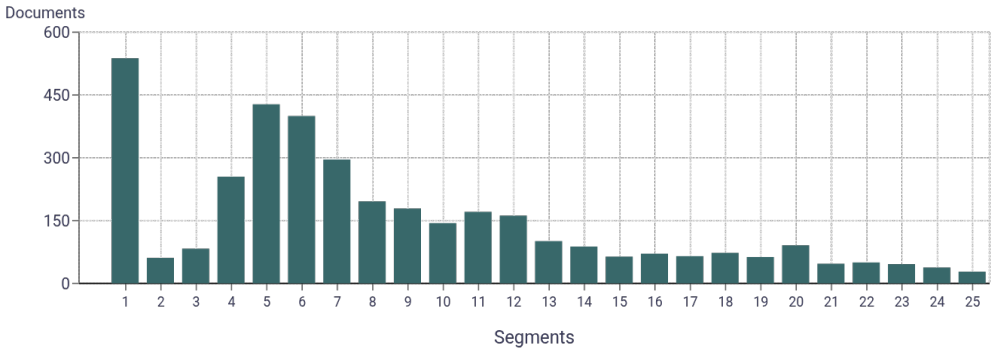
Domain	Docs	% of total
jw.org	1.8K	38.21%
bibles.org	681	14.77%
bible.is	523	11.34%
skyfm.co.ke	519	11.26%
nyimbozakristo.com	146	3.17%
rmsradio.co.ke	137	2.97%
kbc.co.ke	136	2.95%
ebible.org	123	2.67%
blogspot.com	40	0.87%
jaluo.com	35	0.76%

Top 10 TLDs

Domain	Docs	% of total
org	2.7K	58.79%
co.ke	844	18.30%
is	523	11.34%
com	440	9.54%
go.ug	30	0.65%
net	16	0.35%
fr	8	0.17%
org.za	5	0.11%
info	5	0.11%
ru	4	0.09%

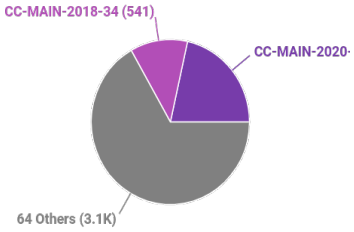
Documents size (in segments)

≤ 25 segments **81.07%** (3.7K documents)
> 25 segments **18.93%** (873 documents)



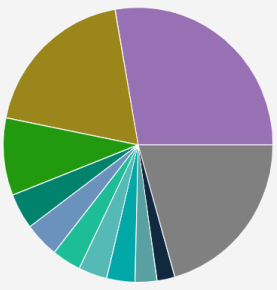
Document collections

CC = 85.36%
IA = 14.64%



Language Distribution

Number of segments in the Luo corpus

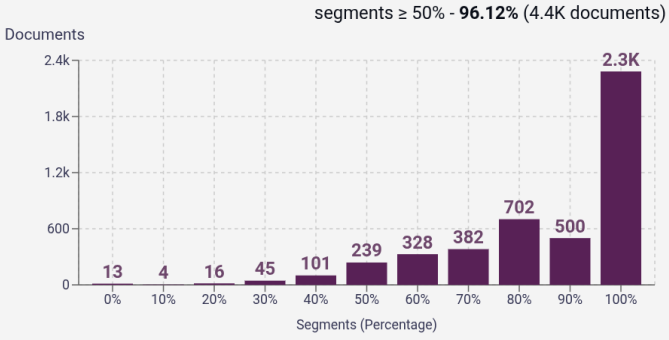


- English - 29K **(27.8%)**
- Esperanto - 20K **(19.1%)**
- Polish - 9.5K **(9.1%)**
- Filipino - 4.4K **(4.2%)**
- Spanish - 4.3K **(4.2%)**
- Italian - 3.6K **(3.5%)**
- Finnish - 3.6K **(3.4%)**
- German - 3.5K **(3.4%)**
- Croatian - 2.7K **(2.6%)**
- Indonesian - 2.2K **(2.1%)**
- 125 Others - 21K **(20.6%)**

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

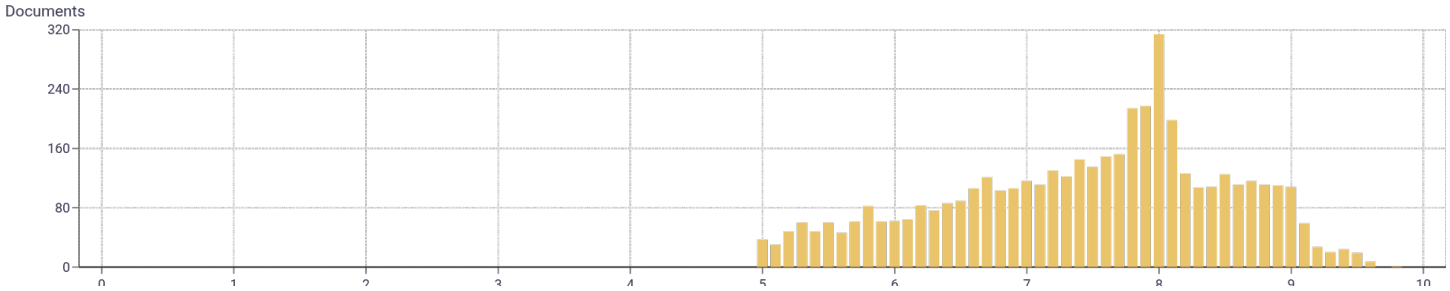
Percentage of segments in Luo inside documents

segments < 50% - **3.88%** (179 documents)
segments ≥ 50% - **96.12%** (4.4K documents)



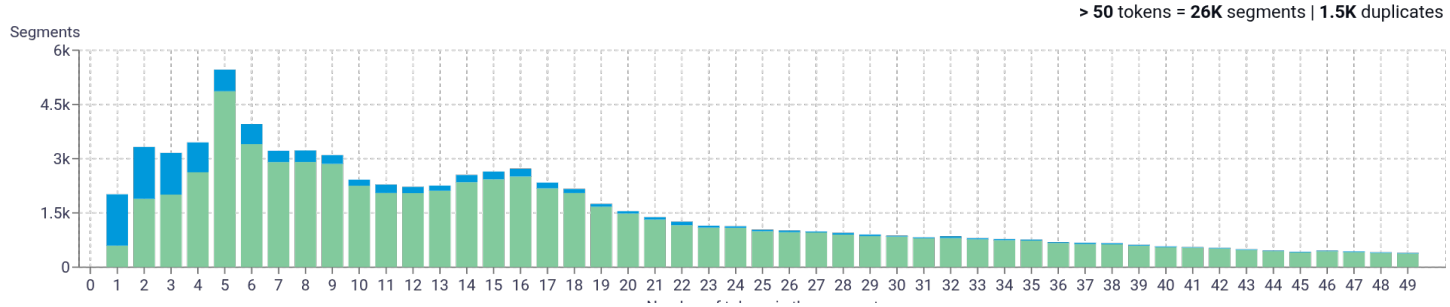
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4.6K documents)

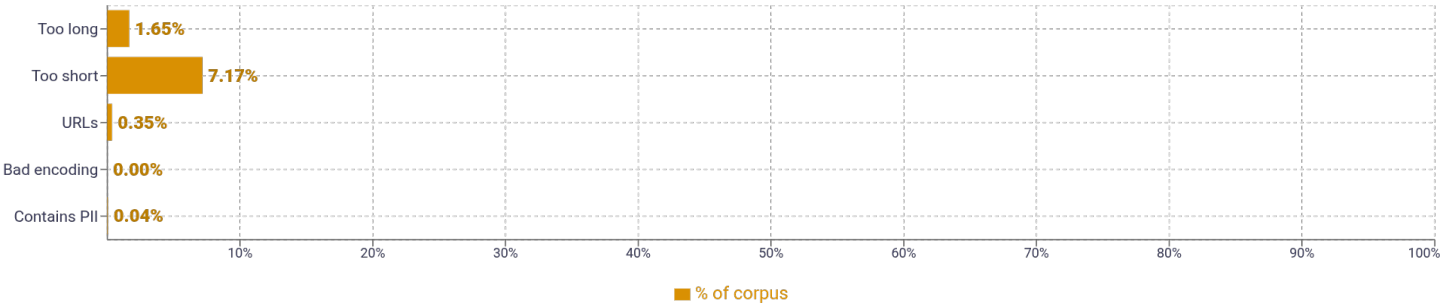


Segment length distribution by token

≤ 49 tokens = **78K** segments | **9.6K** duplicates
> 50 tokens = **26K** segments | **1.5K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ni | 84,586 · i | 76,256 · e | 71,868 · ne | 57,687 · ki | 55,515
2	i kom | 9,043 · ki i | 7,726 · i kare | 4,619 · kata kamano | 3,650 · e piny | 3,617
3	woko ki i | 1,467 · kata obedo ni | 1,362 · gik moko duto | 1,069 · ki i kom | 778 · ato ka ng | 714
4	owaco bot moses ni | 237 · lok me kwena maber | 228 · kwo ma pe tum | 219 · e county ma migori | 205 · gin ma pire tek | 185
5	pi gin ango ma omiyo | 168 · rwot owaco bot moses ni | 115 · gityer ma giwaŋo ki mac | 112 · oko waco ne musa be | 90 · e gima jehova nyasaye wacho | 86

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				