

General overview

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

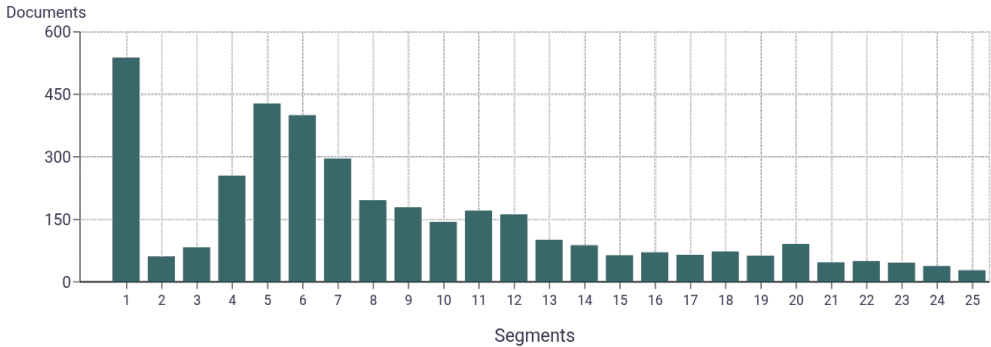
Top 10 domains

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%
jaluocom	35	0.76%

Top 10 TLDs

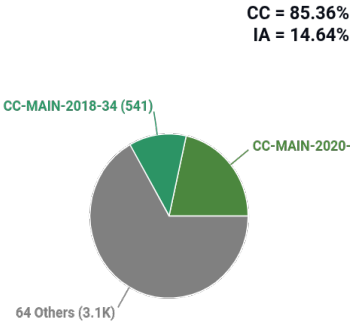
Domain	Docs	% of total
org	2.7K	58.79%
co.ke	844	18.30%
is	523	11.34%
com	440	9.54%
go.ug	30	0.65%
net	16	0.35%
fr	8	0.17%
org.za	5	0.11%
info	5	0.11%
ru	4	0.09%

Documents size (in segments) ⓘ



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

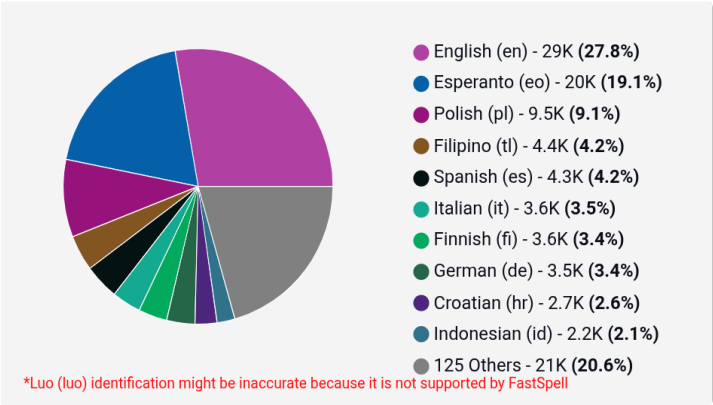
Document collections



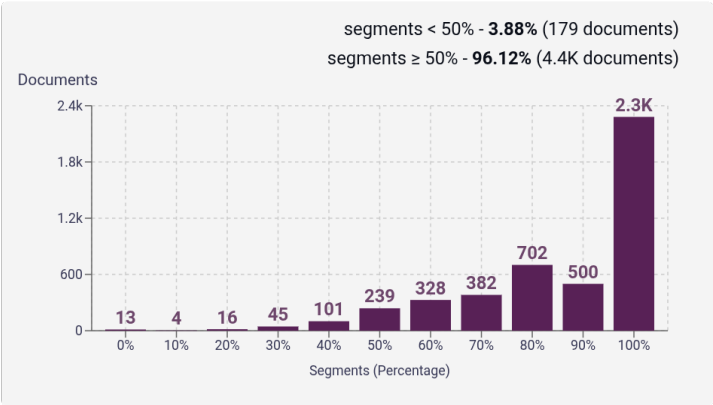
CC = 85.36%
IA = 14.64%

Language Distribution

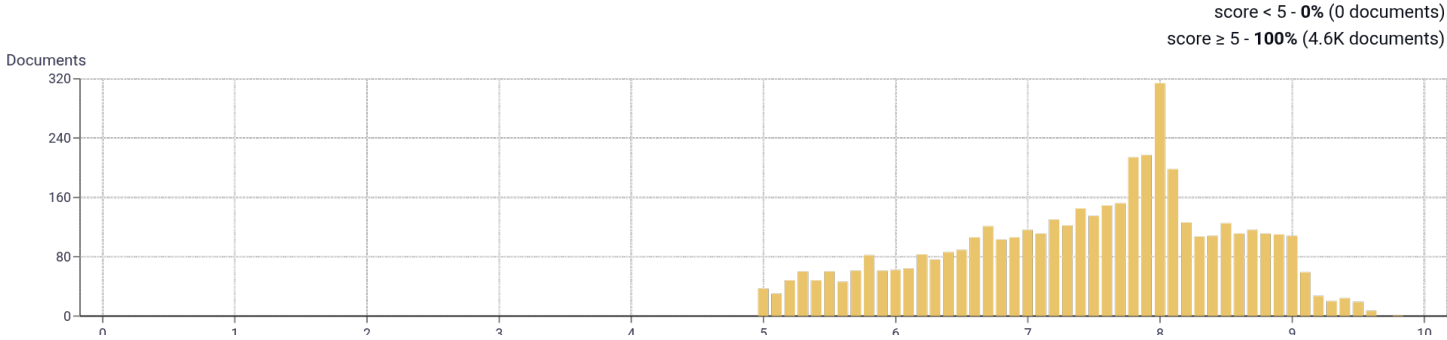
Number of segments in the Luo (luo) corpus



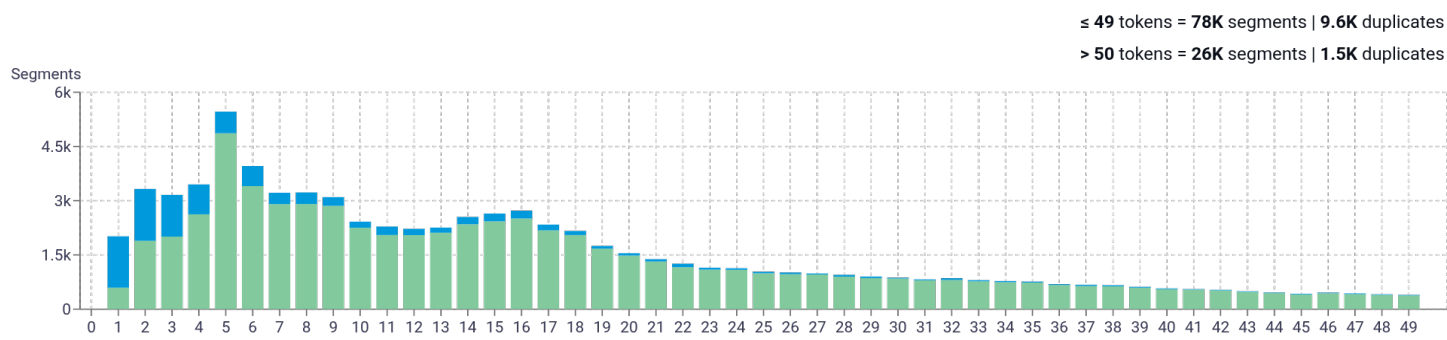
Percentage of segments in Luo (luo) inside documents



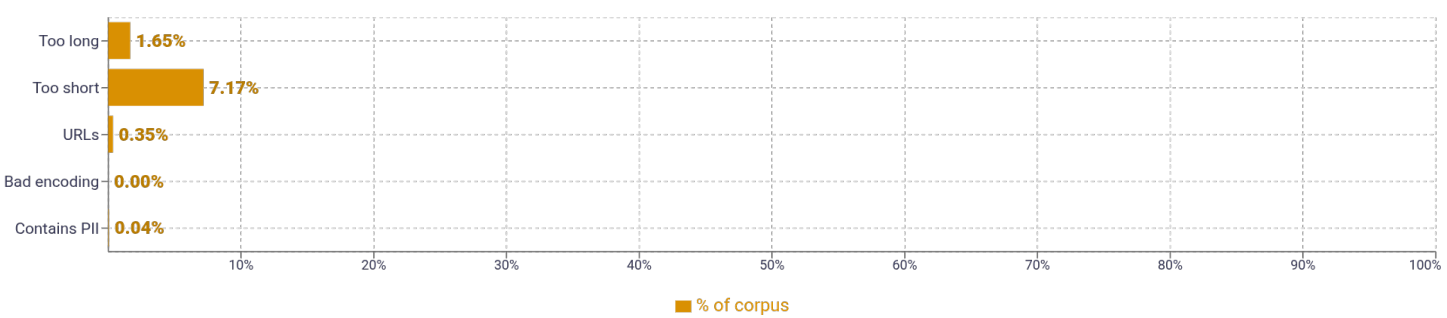
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				