HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-yor_Latn	9/18/2025	Yoruba

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
171,248	3,622,682	2,857,022 (78.86 %)	129M	556,422,071	612.77 MB

Top 10 domains

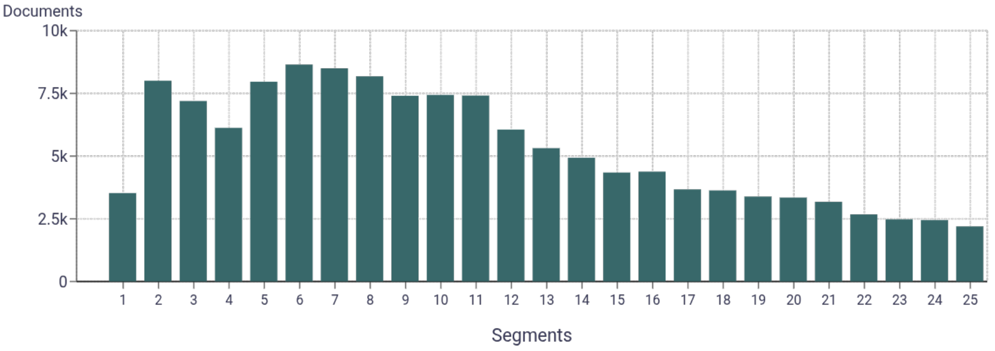
Domain	Docs	% of total
alaroye.org	6.6K	3.84%
ilorin.info	5.8K	3.37%
martech.zone	4.7K	2.74%
awikonko.com.ng	4K	2.31%
androidsis.com	2.8K	1.65%
vessoft.com	2.7K	1.56%
creativosonline...	2.5K	1.43%
bbc.com	2.2K	1.31%
desdelinux.net	2.2K	1.28%
actualidadiphon...	2.1K	1.21%

Top 10 TLDs

Domain	Docs	% of total
com	116K	67.66%
org	16K	9.57%
com.ng	7.2K	4.22%
info	6.3K	3.68%
zone	4.7K	2.74%
net	4.3K	2.50%
tv	1.3K	0.77%
fr	1K	0.60%
gov.ng	914	0.53%
es	881	0.51%

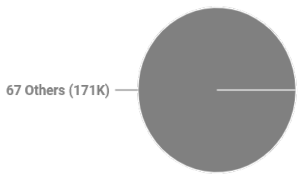
Documents size (in segments) ⓘ

≤ 25 segments **77.34%** (132K documents)
> 25 segments **22.66%** (39K documents)



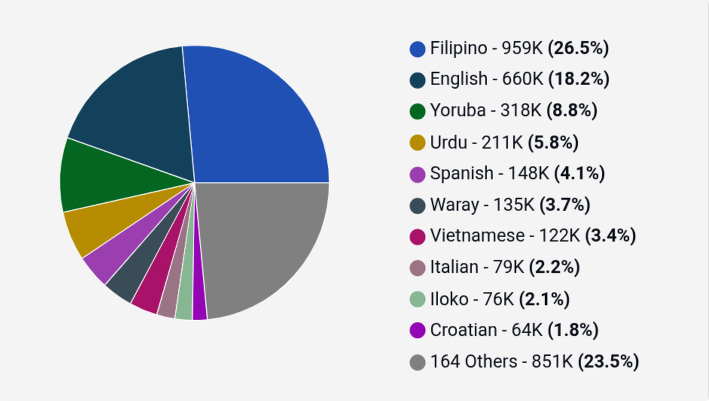
Document collections

CC = 94.45%
IA = 5.55%



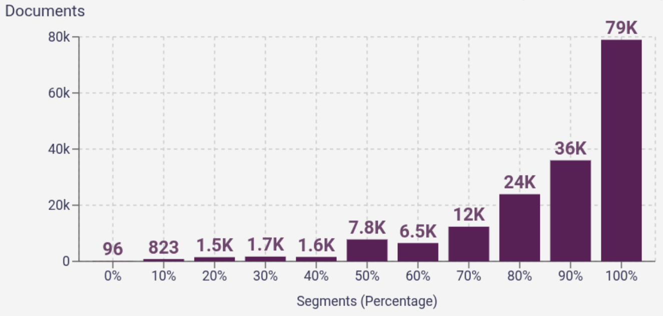
Language Distribution

Number of segments in the Yoruba corpus



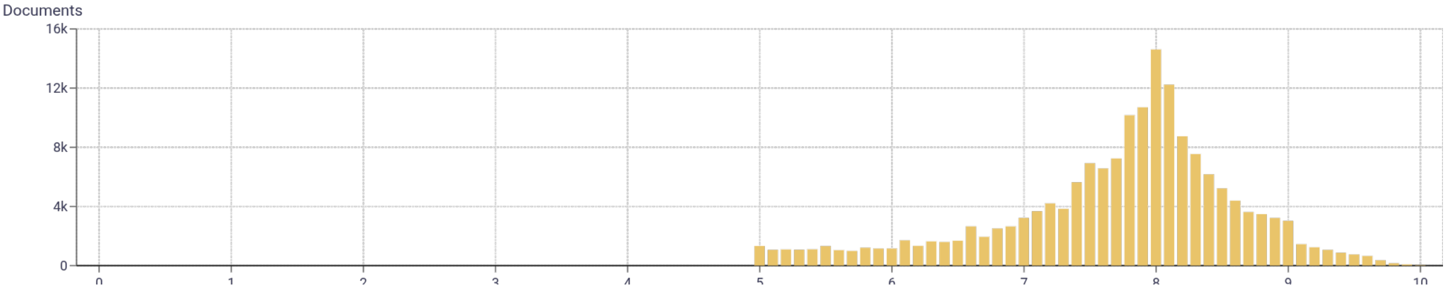
Percentage of segments in Yoruba inside documents

segments < 50% - **3.33%** (5.7K documents)
segments ≥ 50% - **96.67%** (166K documents)



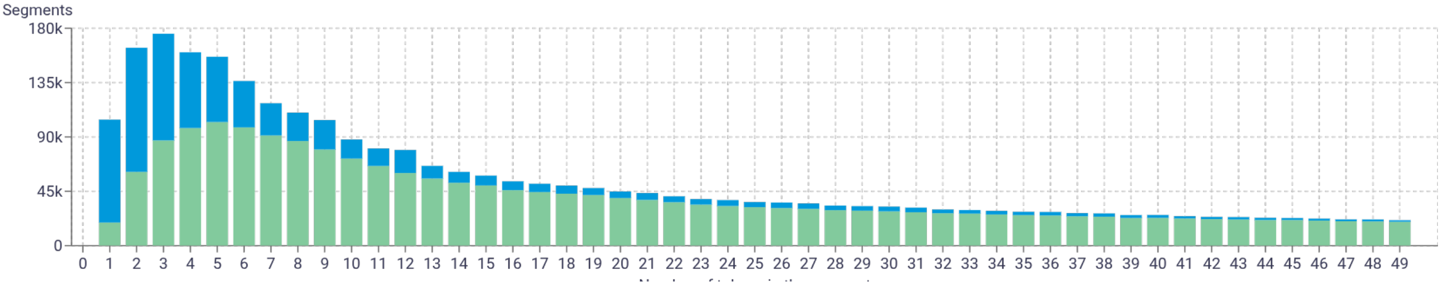
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (171K documents)

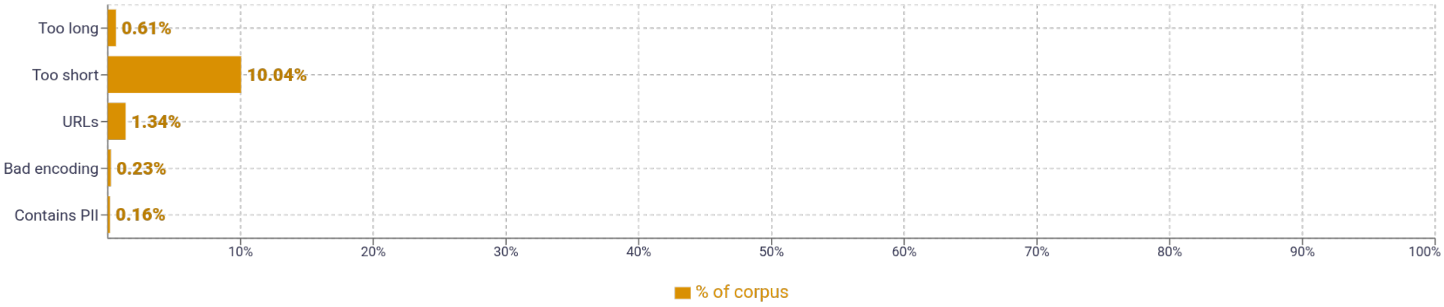


Segment length distribution by token

≤ **49** tokens = **2.8M** segments | **704K** duplicates
> **50** tokens = **846K** segments | **63K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS				
1	wa | 1,144,168	rẹ | 852,782	pe | 816,920	le | 784,495	pẹlu | 752,368
2	ohun elo | 269,116	diẹ sii | 138,632	dara julọ | 124,492	diẹ ninu | 82,837	yẹ ki | 68,612
3	nigbati o ba | 37,615	oju opo wẹẹbu | 36,286	diẹ sii ju | 32,348	rii daju pe | 30,367	pe o le | 22,953
4	bii o ṣe le | 16,516	bi o ṣe le | 15,135	ṣaaju ki o to | 14,519	akọkọ lati sọ ọrọ | 12,823	to ti ni ilọsiwaju | 7,614
5	to read more about it | 5,767	phrase to read more about | 5,767	ọna ti o dara julọ | 3,704	le ṣe iranlọwọ fun ọ | 3,516	ohun ti o dara julọ | 3,494

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				