HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ibo_Latn	9/17/2025	Igbo

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
172,837	4,010,125	3,166,891 (78.97 %)	137M	599,575,524	675.57 MB

Top 10 domains

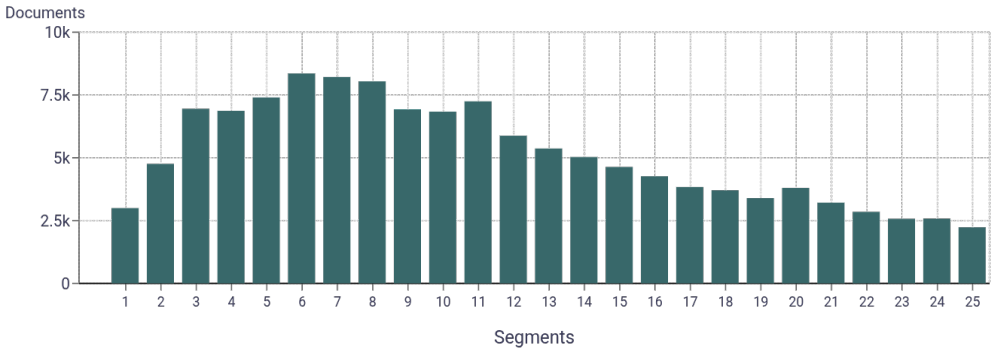
Domain	Docs	% of total
martech.zone	6.9K	3.99%
eturbonews.com	4.5K	2.59%
androidsis.com	4.2K	2.45%
wikipedia.org	3.7K	2.14%
jw.org	3.6K	2.07%
wondershare.com	2.4K	1.41%
von.gov.ng	2.3K	1.32%
wikiquote.org	1.9K	1.11%
kaoditaa.com	1.9K	1.09%
actualidadgadge...	1.9K	1.09%

Top 10 TLDs

Domain	Docs	% of total
com	129K	74.44%
org	15K	8.73%
zone	6.9K	3.99%
net	4.5K	2.61%
gov.ng	2.3K	1.32%
ru	1.8K	1.04%
com.ng	1.2K	0.69%
fr	1.2K	0.68%
st	682	0.39%
es	588	0.34%

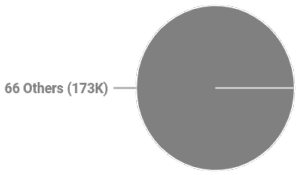
Documents size (in segments)

≤ 25 segments **74.08%** (128K documents)
> 25 segments **25.92%** (45K documents)



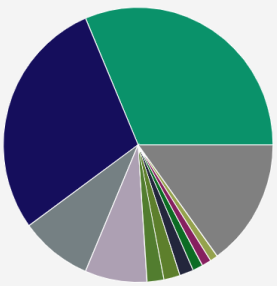
Document collections

CC = 95.25%
IA = 4.75%



Language Distribution

Number of segments in the Igbo corpus

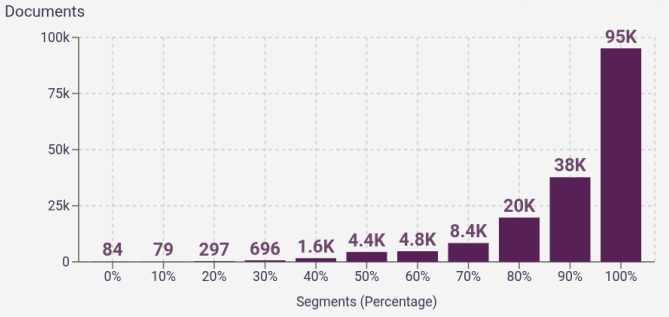


*Igbo identification might be inaccurate because it is not supported by FastSpell.

- Vietnamese - 1.3M **(31.4%)**
- English - 1.1M **(28.6%)**
- Croatian - 345K **(8.6%)**
- Filipino - 298K **(7.4%)**
- French - 82K **(2.0%)**
- Italian - 78K **(1.9%)**
- German - 67K **(1.7%)**
- Swahili - 49K **(1.2%)**
- Spanish - 45K **(1.1%)**
- Haitian Creole - 38K **(0.9%)**
- 163 Others - 602K **(15.0%)**

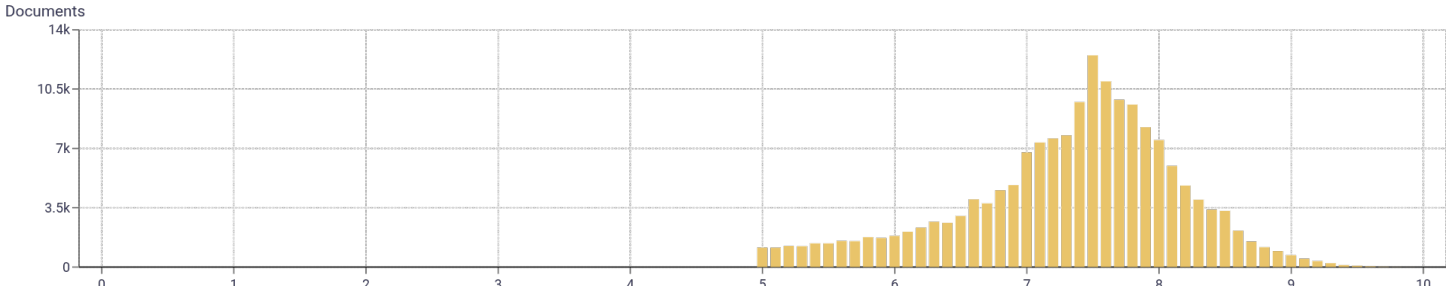
Percentage of segments in Igbo inside documents

segments < 50% - **1.62%** (2.8K documents)
segments ≥ 50% - **98.38%** (170K documents)



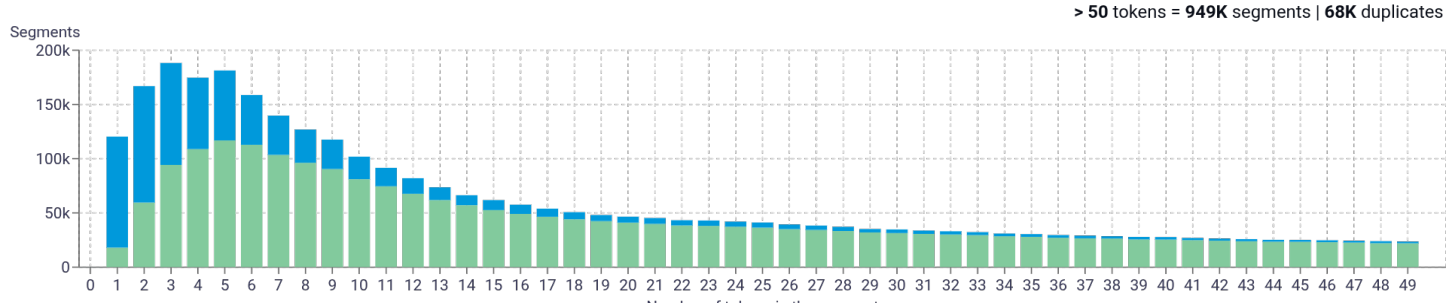
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (173K documents)

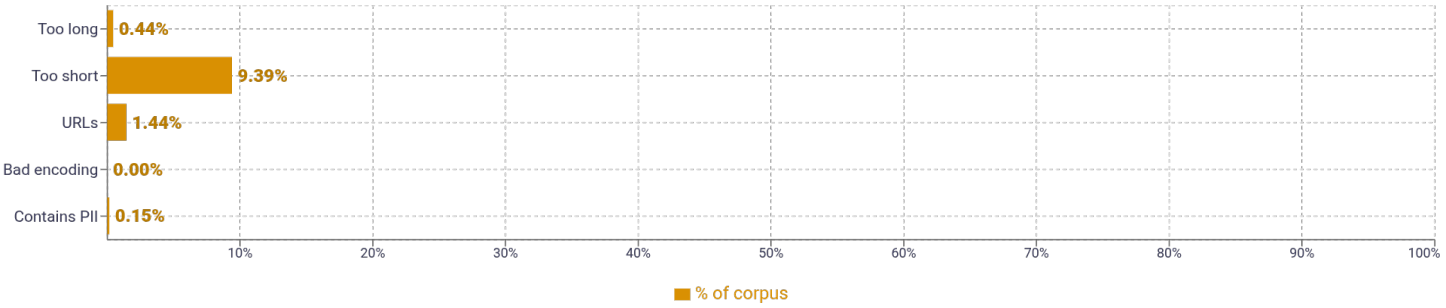


Segment length distribution by token

≤ 49 tokens = **3.1M** segments | **777K** duplicates
> 50 tokens = **949K** segments | **68K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	na- | 3,892,065 · n' | 2,308,777 · ike | 1,266,406 · ahụ | 1,145,658 · ọrụ | 1,060,753	
2	nwere ike | 544,280 · ụlọ ọrụ | 249,650 · na- eme | 241,374 · iche iche | 195,792 · n' ihi | 185,509	
3	na- arụ ọrụ | 95,955 · anyị nwere ike | 49,576 · onye ọ bụla | 47,367 · ole na ole | 39,783 · aga n' ihu | 39,540	
4	na- aga n' ihu | 32,514 · bụrụ na ị na- | 24,337 · akụkọ ihe mere eme | 20,411 · nta ka ọ bụrụ | 15,378 · igwe anaghị agba nchara | 14,337	
5	bụrụ onye mbụ ịza ajụjụ | 11,755 · na- aga n' ihu na- | 8,690 · na- arụ ọrụ nke ọma | 8,504 · na- eme ka ọ bụrụ | 7,397 · n' oge na- adighị anya | 6,351	

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				