

General overview

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

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

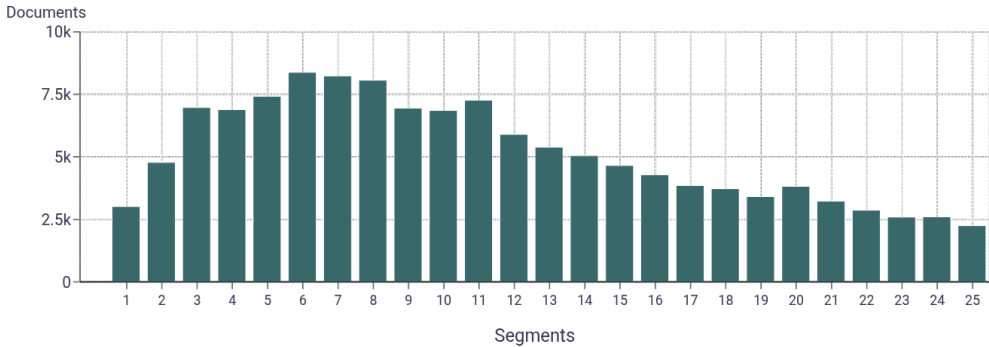
Top 10 domains

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%
actualidadgadgets.com	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%

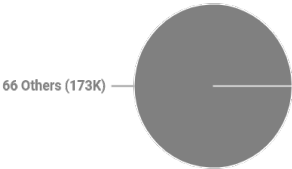
Documents size (in segments) ⓘ



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

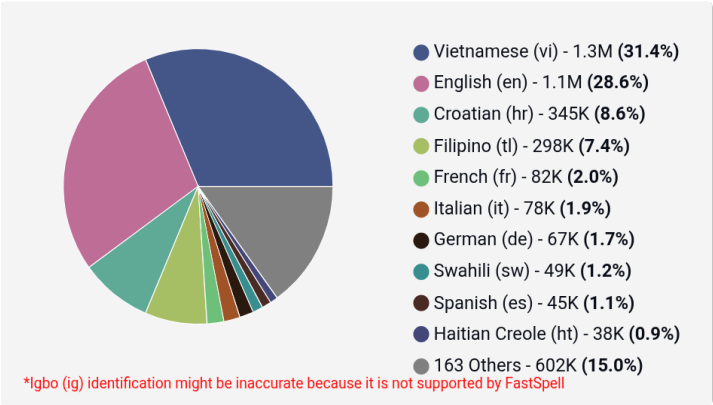
Document collections

CC = 95.25%
IA = 4.75%

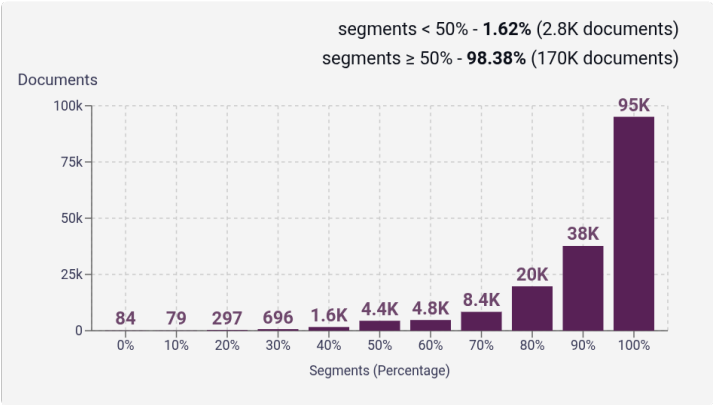


Language Distribution

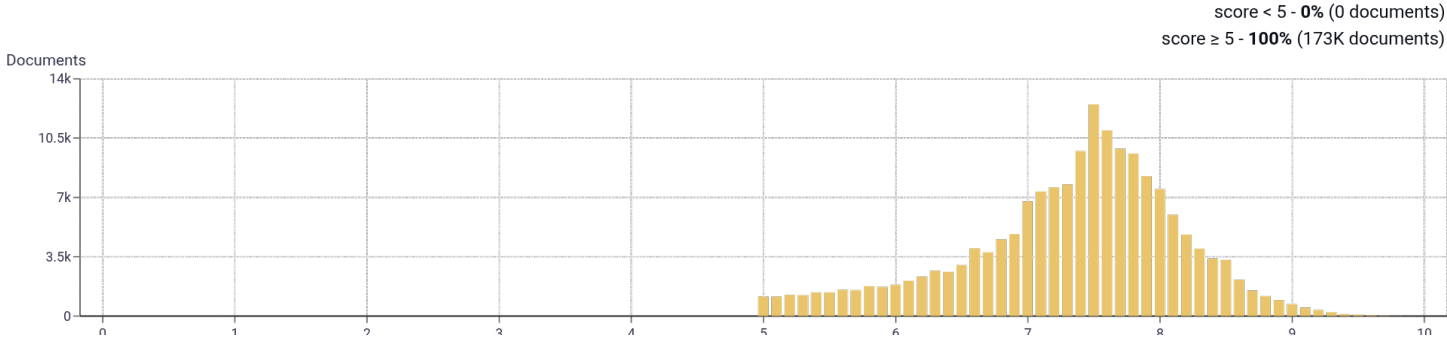
Number of segments in the Igbo (ig) corpus



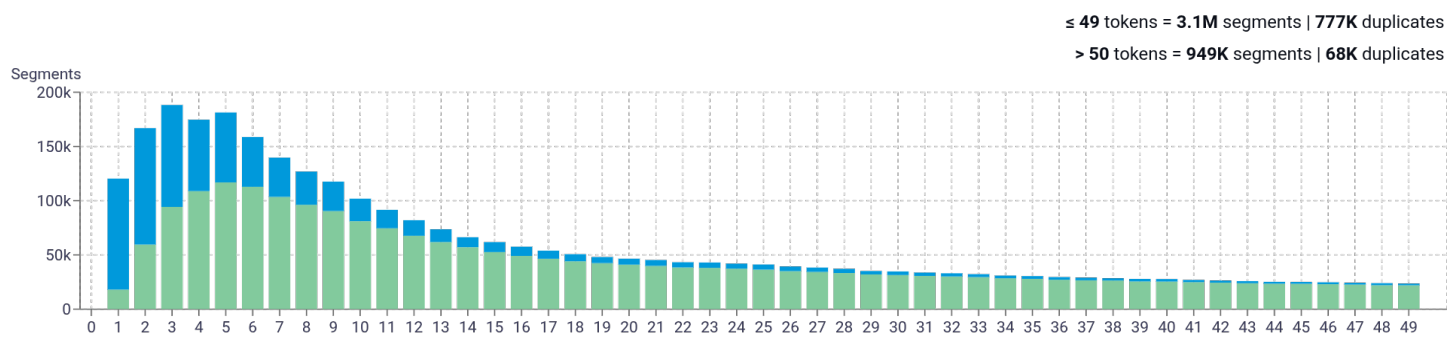
Percentage of segments in Igbo (ig) inside documents



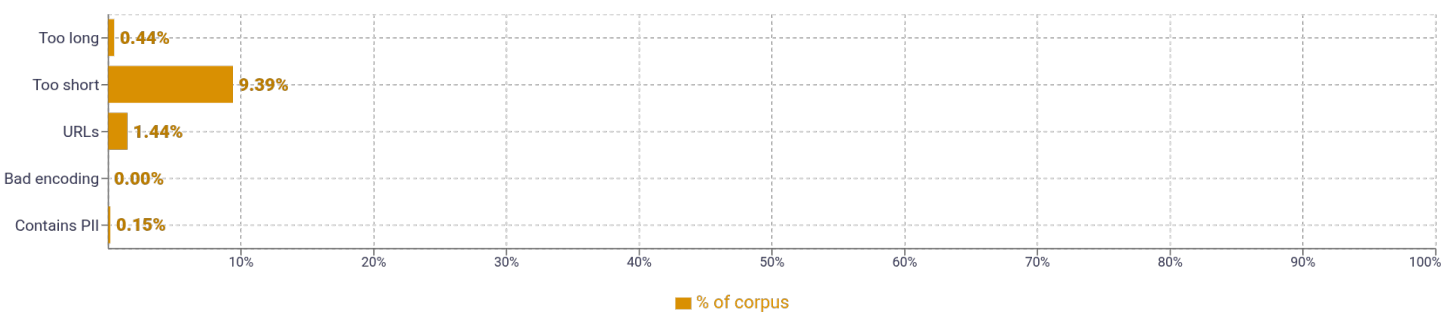
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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				