

General overview

Corpus	Date	Language
hplt-v3-fuv_Latn	9/17/2025	Nigerian Fulfulde (fuv)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
9,972	193,029	154,507 (80.04 %)	7.2M	34,767,218	34.74 MB

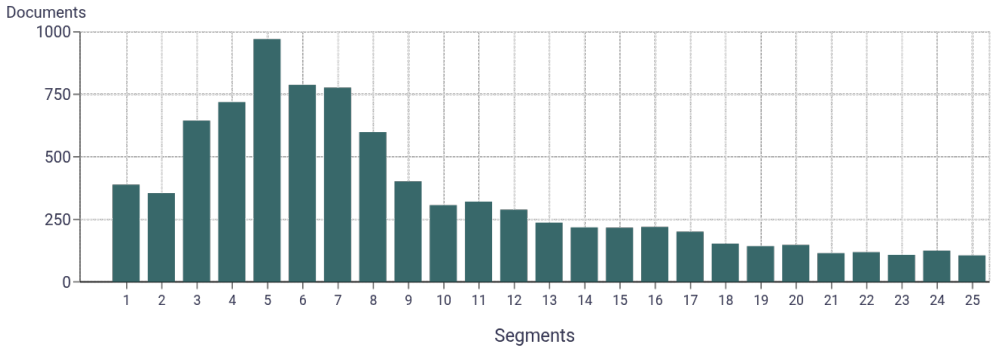
Top 10 domains

Domain	Docs	% of total
pulaar.org	2.4K	24.10%
von.gov.ng	1.6K	16.33%
rfi.fr	684	6.86%
wikipedia.org	619	6.21%
ebible.org	487	4.88%
fuutamedia.com	444	4.45%
dingiralfulbe.com	356	3.57%
bible.is	346	3.47%
binndipulaar.com	317	3.18%
pulaaronline.com	269	2.70%

Top 10 TLDs

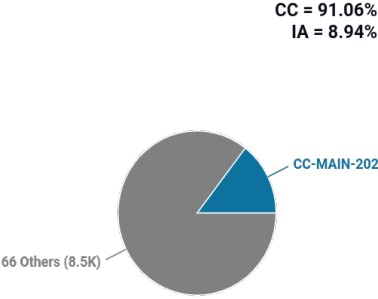
Domain	Docs	% of total
org	3.9K	39.36%
com	2.3K	22.64%
gov.ng	1.6K	16.33%
net	816	8.18%
fr	698	7.00%
is	346	3.47%
ir	120	1.20%
info	63	0.63%
eus	56	0.56%
sn	7	0.07%

Documents size (in segments) ⓘ



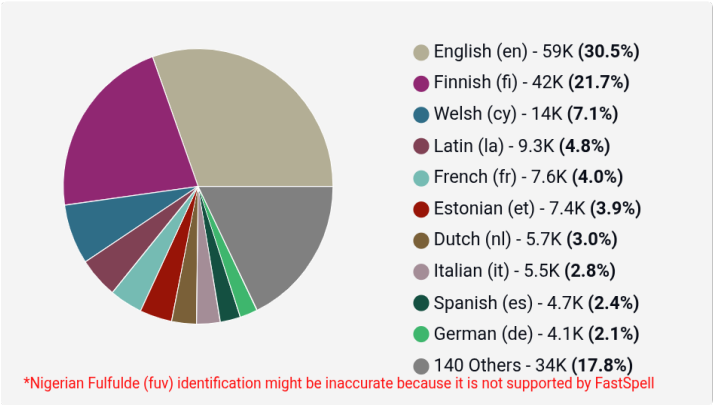
≤ 25 segments **86.96%** (8.7K documents)
> 25 segments **13.04%** (1.3K documents)

Document collections

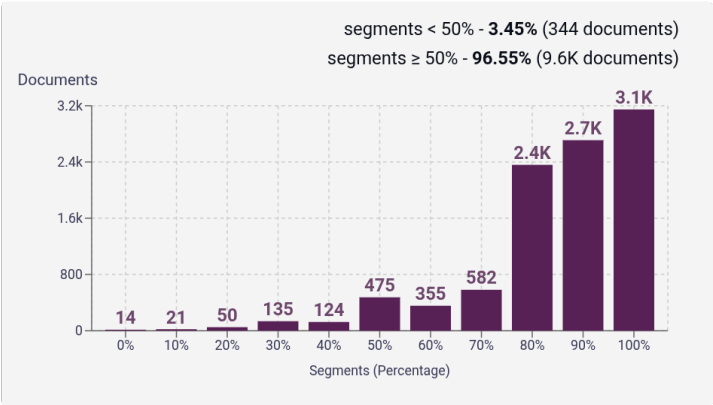


Language Distribution

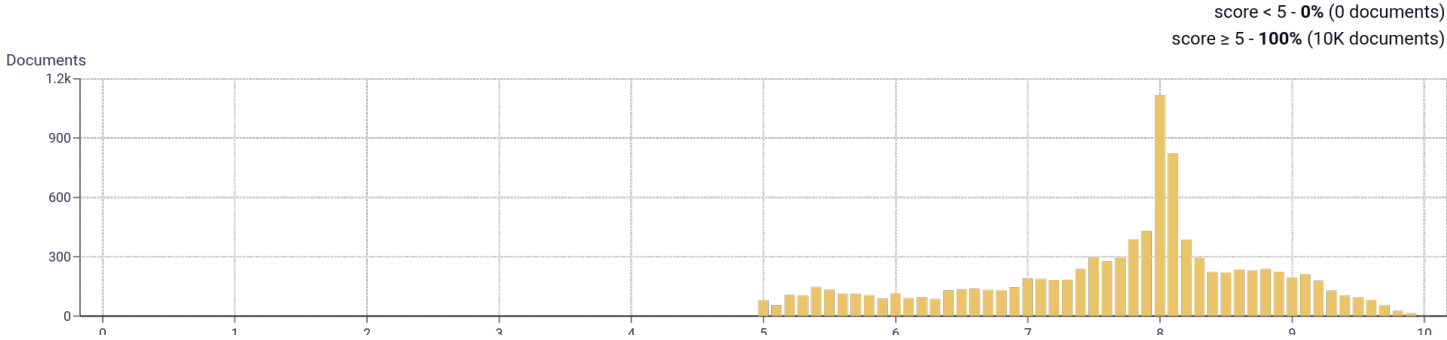
Number of segments in the Nigerian Fulfulde (fuv) corpus



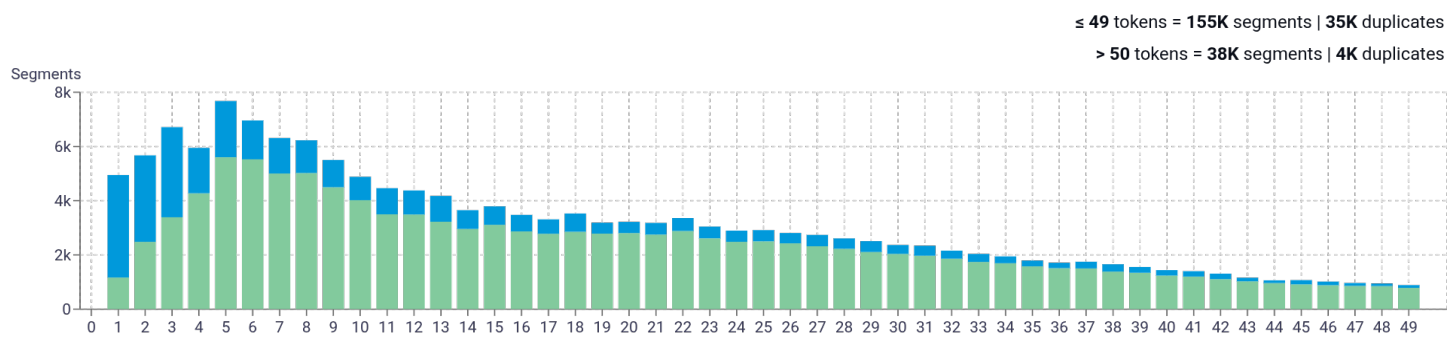
Percentage of segments in Nigerian Fulfulde (fuv) inside documents



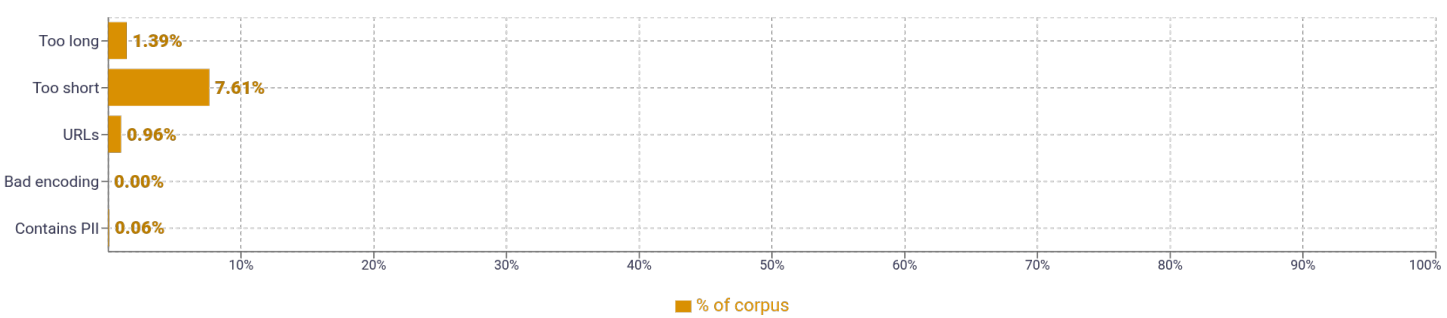
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	nder 42,686 mum 34,084 ina 31,190 makko 29,075 leydi 22,785	
2	so tawii 3,310 leydi ndii 3,089 lesdi naajeeriya 2,683 he nder 2,659 si tawii 2,326	
3	hay so tawii 777 tayto dadi wiki 586 fedde bambaare pulaar 586 alaa e sago 584 bookara aamadu bah 568	
4	yah yah yah yah 455 bambaare pulaar e muritani 249 suwaa tawo wadeede doo 180 so tawii ada anndi 180 amen facebook e twitter 177	
5	yah yah yah yah yah 423 fedde bambaare pulaar e muritani 248 kelle amen facebook e twitter 172 fow ka kelle amen facebook 161 de boghe menace la population 158	

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				