

General overview

Corpus	Date	Language
hplt-v3-ktu_Latn	9/17/2025	Kongo (kg)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,423	86,549	79,309 (91.63 %)	4.7M	22,319,478	21.48 MB

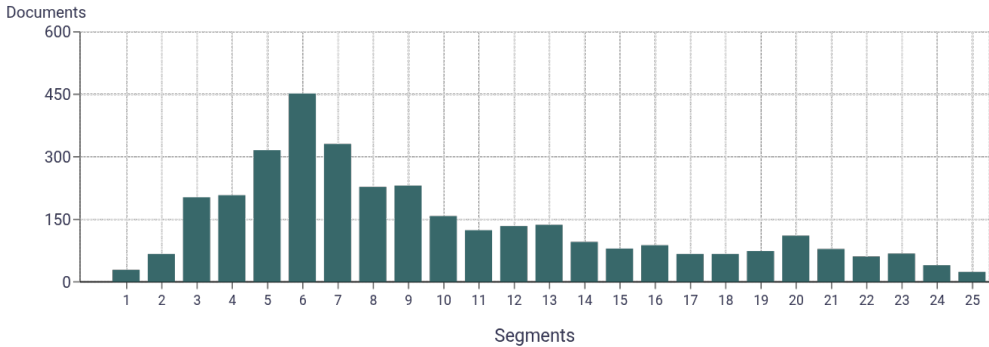
Top 10 domains

Domain	Docs	% of total
jw.org	3.8K	85.12%
radiookapi.net	294	6.65%
wikipedia.org	75	1.70%
une.cd	41	0.93%
bibles.org	30	0.68%
editorial7.net	21	0.47%
biblafrique.org	19	0.43%
grindr.com	18	0.41%
biblafrique.net	13	0.29%
gotquestions.org	10	0.23%

Top 10 TLDs

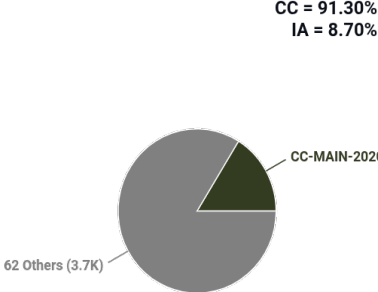
Domain	Docs	% of total
org	3.9K	89.22%
net	344	7.78%
com	75	1.70%
cd	41	0.93%
club	6	0.14%
ru	2	0.05%
click	2	0.05%
pt	1	0.02%
io	1	0.02%
info	1	0.02%

Documents size (in segments) ⓘ



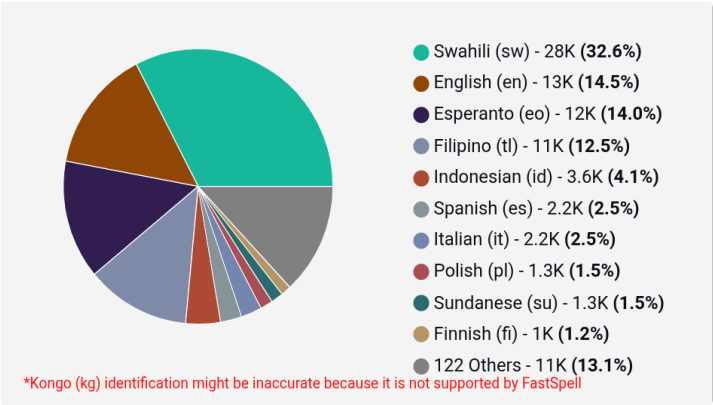
≤ 25 segments **78.52%** (3.5K documents)
> 25 segments **21.48%** (950 documents)

Document collections

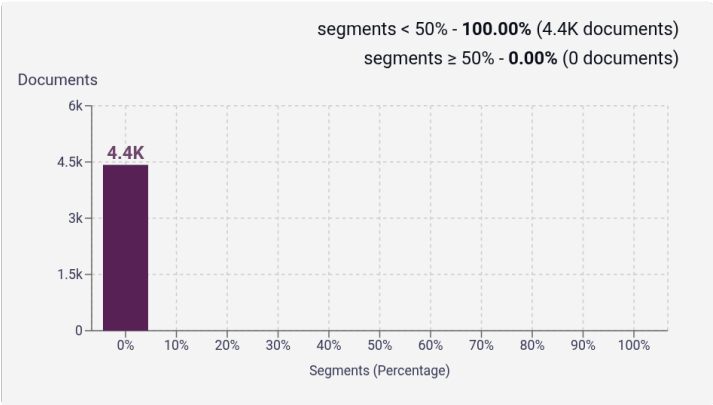


Language Distribution

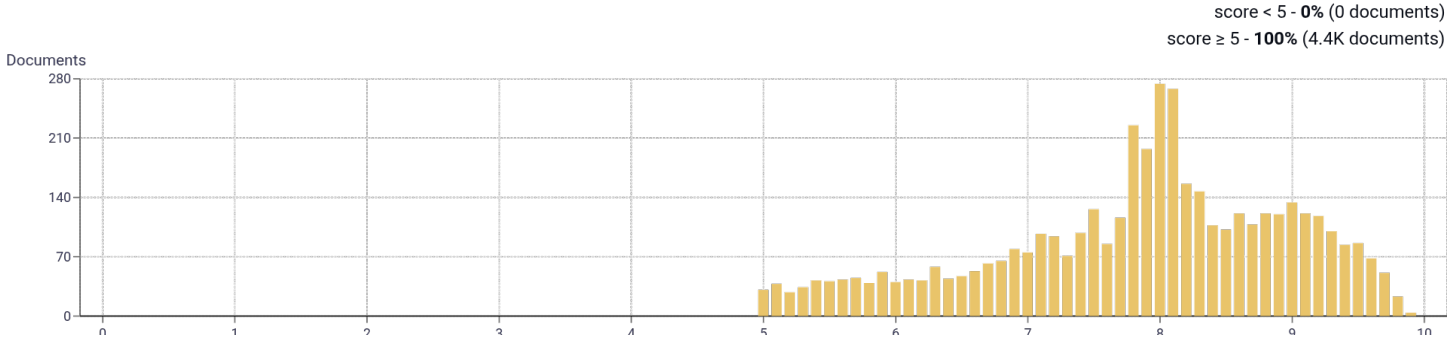
Number of segments in the Kongo (kg) corpus



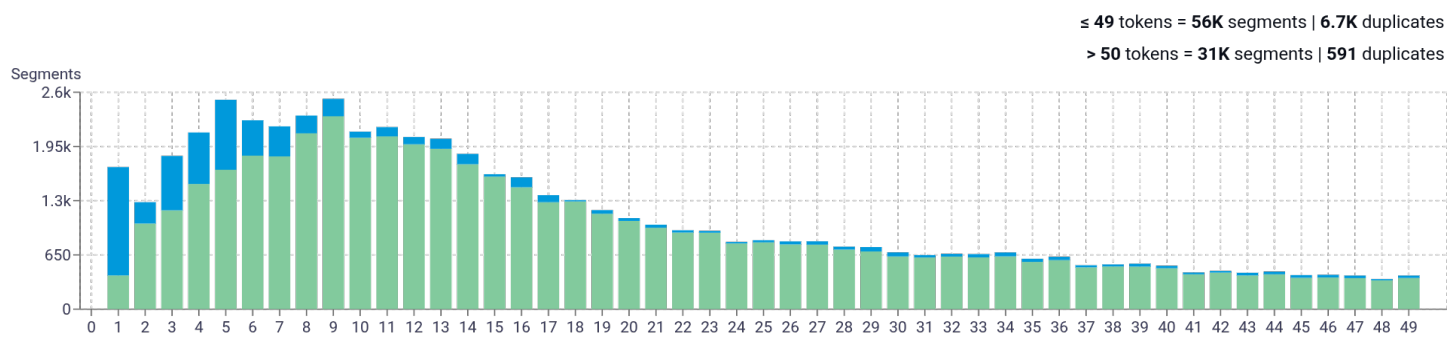
Percentage of segments in Kongo (kg) inside documents



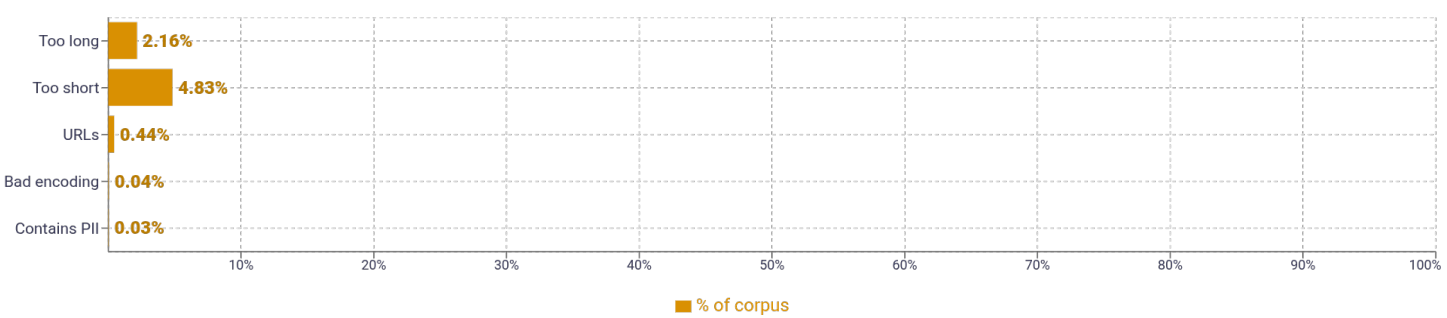
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ve 36,620bantu 34,673nzambi 31,738mambu 28,186mutindu 21,102	
2	diaka ve 1,402kuma kia 1,245mutindu yehowa 838mpeve santu 821wavova kwa 770	
3	bantu ya nkaka 2,490bambangi ya yehowa 1,422bantu ya izraele 1,188ndinga ya nzambi 1,112ntoto ya mvimba 1,106	
4	mambu yina ta salama 222bantu ya kukonda kukuka 204kulonga nsangu ya mbote 191zinga mutindu bakristu fwete 168mutindu bakristu fwete zinga 168	
5	luzingu ya mvula na mvula 417bakristu yina bo me tulaka 314bimvwama ya ndinga ya nzambi 238lutangu ya biblia ya mposho 204zinga mutindu bakristu fwete zinga 167	

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				