HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lin_Latn	9/18/2025	Lingala

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
13,561	441,493	395,334 (89.54 %)	16M	80,899,196	78.36 MB

Top 10 domains

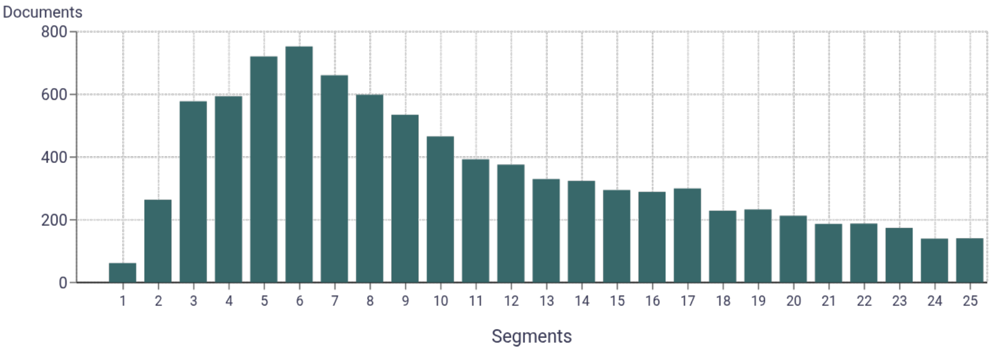
Domain	Docs	% of total
jw.org	2.7K	19.98%
congomikili.com	2.2K	16.16%
voalingala.com	2.2K	15.99%
ebible.org	525	3.87%
voiceofcongo.net	489	3.61%
mbokamosika.com	308	2.27%
wikipedia.org	297	2.19%
rdcongoinfos.com	285	2.10%
radiookapi.net	251	1.85%
skyrock.com	237	1.75%

Top 10 TLDs

Domain	Docs	% of total
com	7K	51.66%
org	4.1K	30.03%
net	1.2K	8.70%
info	387	2.85%
fr	118	0.87%
ch	94	0.69%
cat	87	0.64%
ru	79	0.58%
me	47	0.35%
eu	47	0.35%

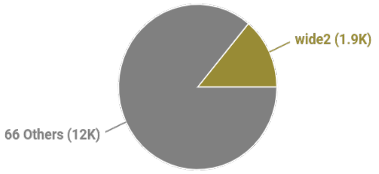
Documents size (in segments) ⓘ

≤ **25** segments **66.7%** (9K documents)
> **25** segments **33.3%** (4.5K documents)



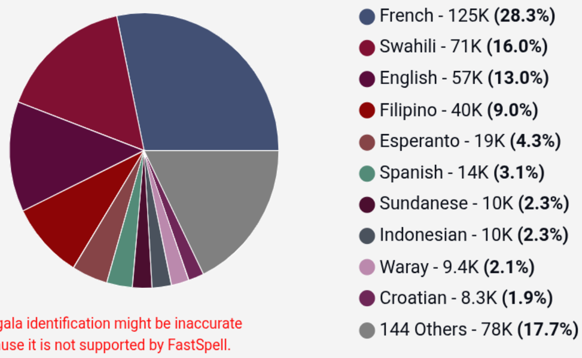
Document collections

CC = 75.33%
IA = 24.67%



Language Distribution

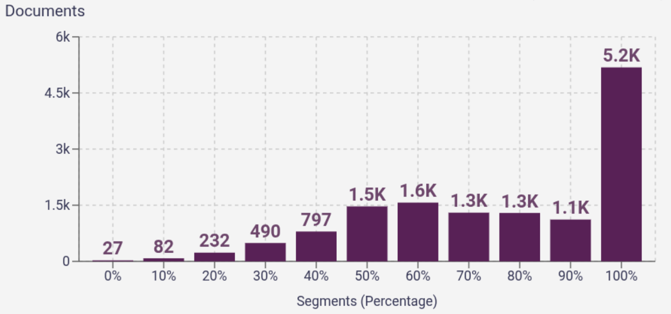
Number of segments in the Lingala corpus



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

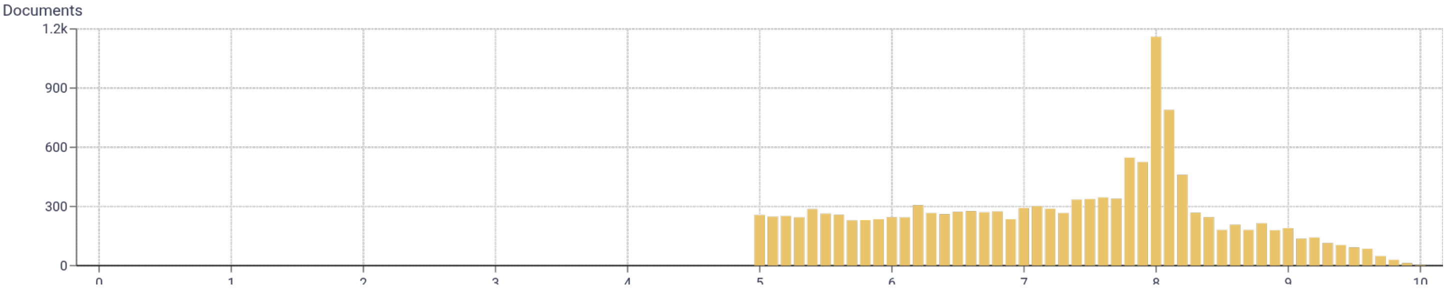
Percentage of segments in Lingala inside documents

segments < 50% - **12.01%** (1.6K documents)
segments ≥ 50% - **87.99%** (12K documents)



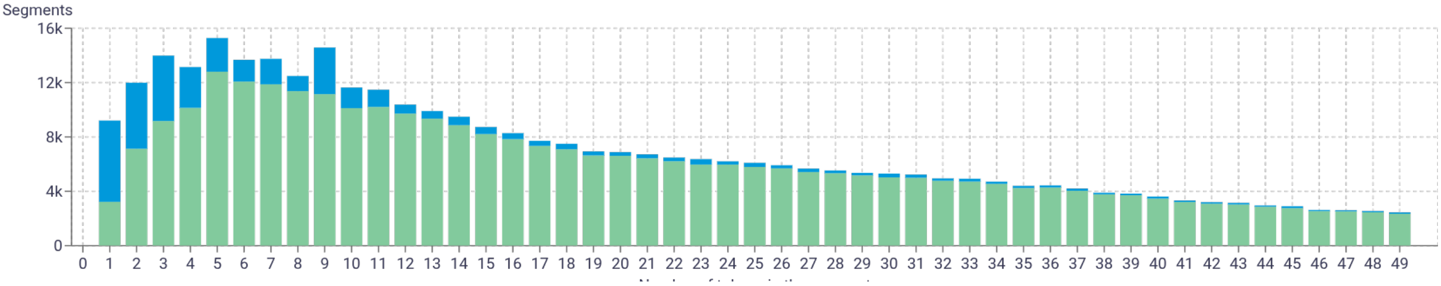
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (14K documents)

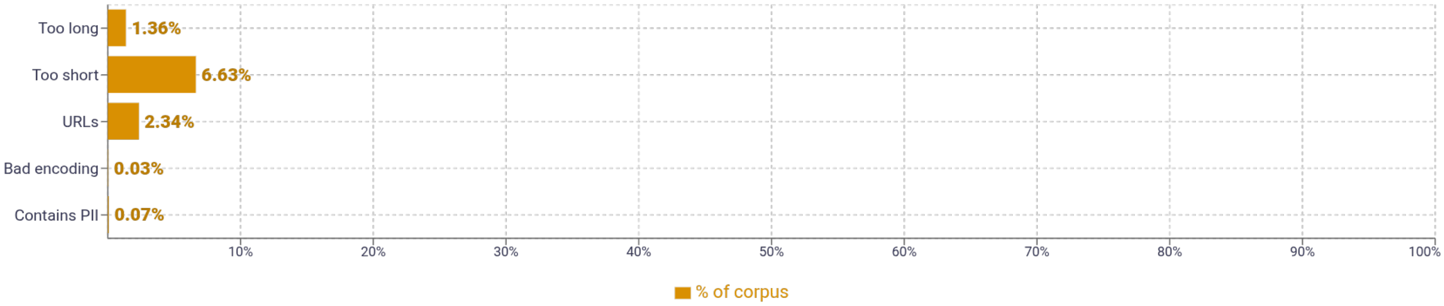


Segment length distribution by token

≤ **49** tokens = **349K** segments | **41K** duplicates
> **50** tokens = **93K** segments | **4.9K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ba | 242,704 oyo | 206,345 ye | 135,707 de | 119,380 yo | 114,199
2	de la | 14,223 oyo ezali | 12,254 il faut | 11,027 de l | 8,730 ba congolais | 8,038
3	kie kie kie | 3,302 est ce que | 2,371 batatoli ya yehova | 1,722 tout le monde | 1,718 awa na poto | 1,714
4	kie kie kie kie | 1,948 lire tous les commentaires | 1,681 pour lire tous les | 1,679 sur anciens commentaires pour | 1,678 commentaires pour lire tous | 1,678
5	pour lire tous les commentaires | 1,679 sur anciens commentaires pour lire | 1,678 commentaires pour lire tous les | 1,678 appuyez sur anciens commentaires pour | 1,678 anciens commentaires pour lire tous | 1,678

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				