HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bam_Latn	9/17/2025	Bambara (bm)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,638	64,632	51,831 (80.19 %)	2.6M	11,222,879	11.52 MB

Top 10 domains

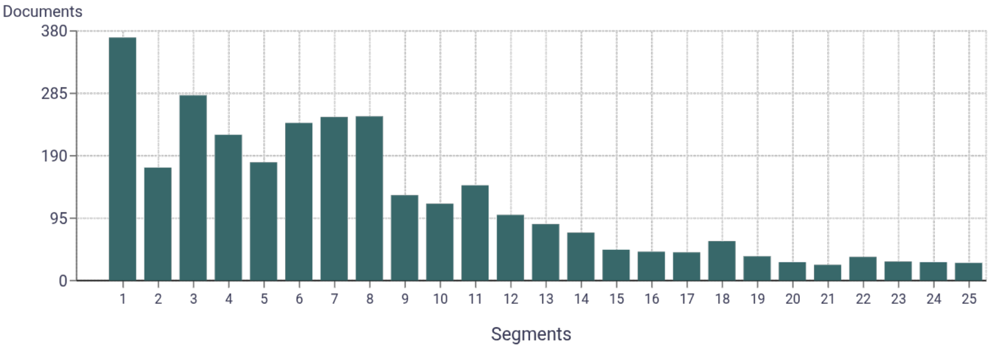
Domain	Docs	% of total
fakan.ml	457	12.56%
rfi.fr	386	10.61%
dokotoro.org	297	8.16%
bible.is	259	7.12%
voabambara.com	196	5.39%
wikipedia.org	179	4.92%
donkibaru.ml	167	4.59%
islamhouse.com	161	4.43%
thieme.com	128	3.52%
jw.org	116	3.19%

Top 10 TLDs

Domain	Docs	% of total
com	1.1K	30.10%
org	697	19.16%
ml	644	17.70%
fr	414	11.38%
is	259	7.12%
net	163	4.48%
gov	81	2.23%
ir	63	1.73%
ru	47	1.29%
co	41	1.13%

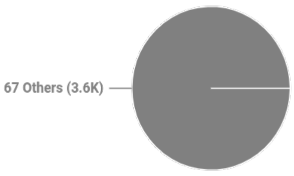
Documents size (in segments) ⓘ

≤ 25 segments **82.99%** (3K documents)
> 25 segments **17.01%** (619 documents)



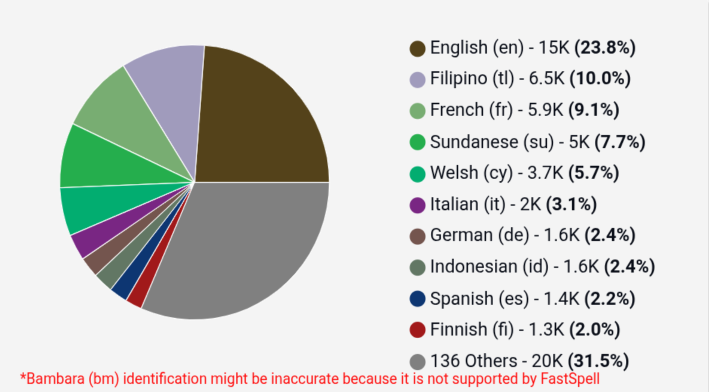
Document collections

CC = 91.51%
IA = 8.49%



Language Distribution

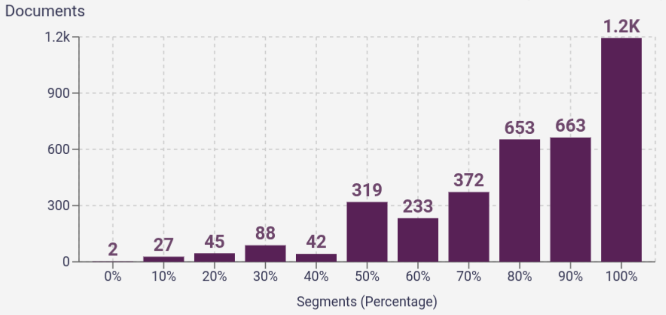
Number of segments in the Bambara (bm) corpus



*Bambara (bm) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Bambara (bm) inside documents

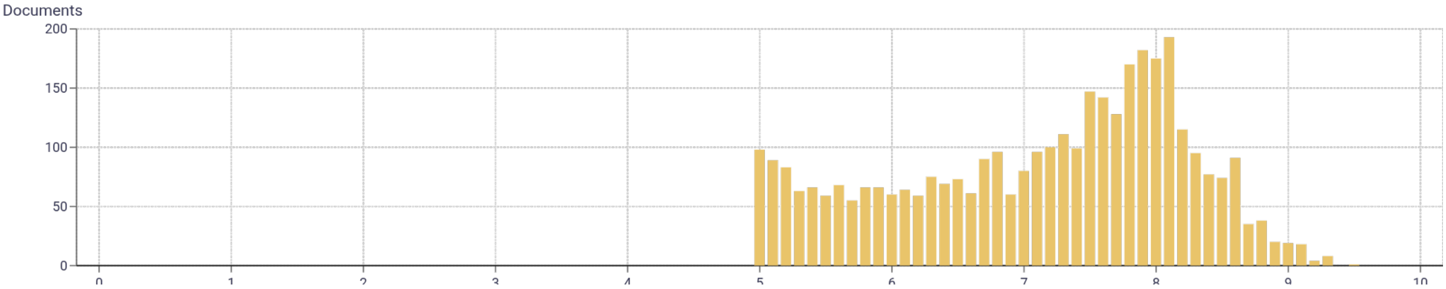
segments < 50% - **5.61%** (204 documents)
segments ≥ 50% - **94.39%** (3.4K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

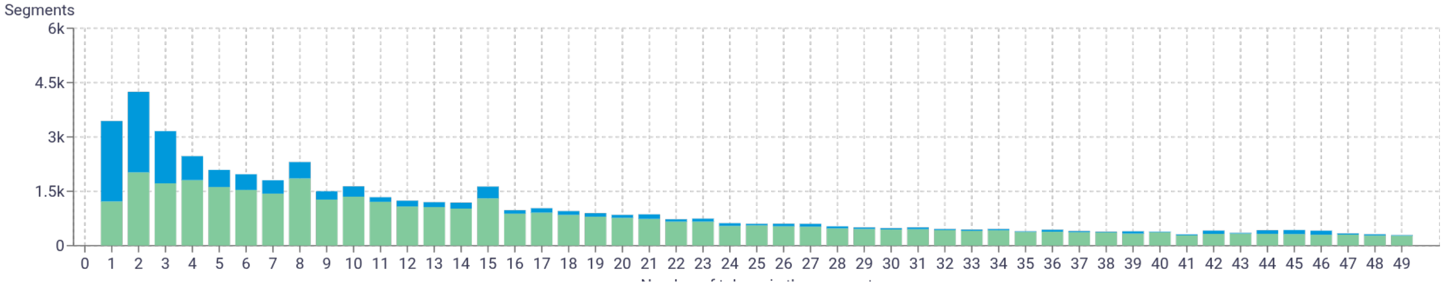
score ≥ 5 - **100%** (3.6K documents)



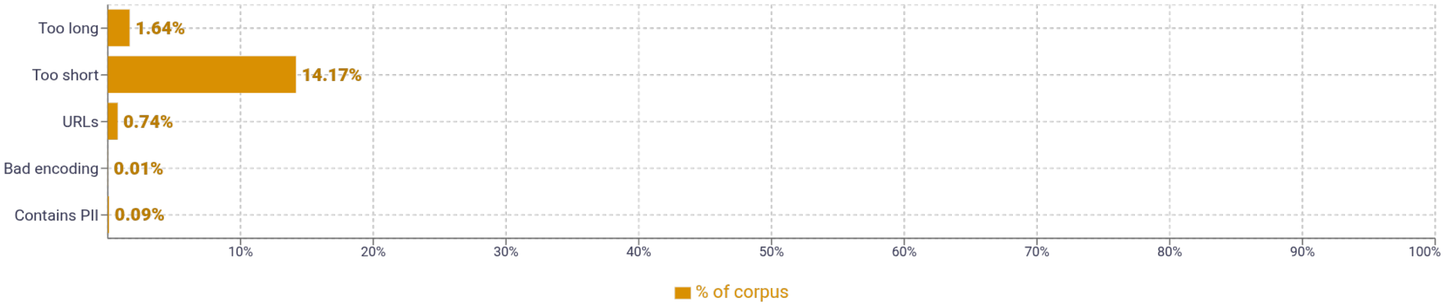
Segment length distribution by token

≤ 49 tokens = **50K** segments | **12K** duplicates

> 50 tokens = **15K** segments | **1.1K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ee | 27,294 eee | 25,379 na | 22,468 eeee | 18,078 k | 14,224
2	ee eee | 5,449 eeeeeeee ee | 3,254 eeeeeeeee ee | 2,736 eee eeeeeeee | 2,654 eee eeee | 2,595
3	ee eee eeeeeeee | 860 eee eeeeeeee ee | 830 eee eeeeeeeee ee | 779 cogo min na | 712 eeeeeeee ee eee | 698
4	ee eee eeeeeeee ee | 343 jamanakuntigi koloneli asimi goyita | 266 access the full content | 249 ee eee eeeeeeeee ee | 239 eee eeeeeeeee ee eee | 151
5	up for a free trial | 147 trial to access the full | 147 sign up for a free | 147 please login or sign up | 147 login or sign up for | 147

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				