HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-mos_Latn	9/18/2025	Mossi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,892	47,988	43,408 (90.46 %)	3.3M	11,592,273	12.53 MB

Top 10 domains

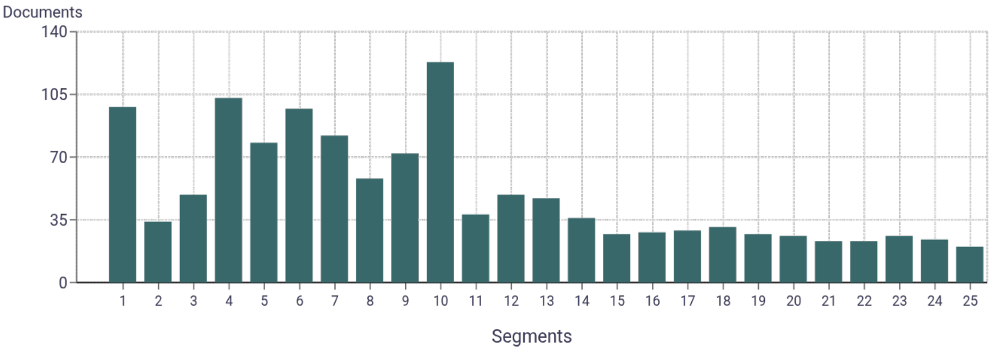
Domain	Docs	% of total
jw.org	1.3K	66.49%
wikimedia.org	191	10.10%
raamde-bf.com	181	9.57%
bible.is	95	5.02%
islamhouse.com	28	1.48%
wikipedia.org	19	1.00%
player.fm	17	0.90%
bible.com	15	0.79%
hadeethenc.com	11	0.58%
islamenc.com	9	0.48%

Top 10 TLDs

Domain	Docs	% of total
org	1.5K	78.86%
com	268	14.16%
is	95	5.02%
fm	17	0.90%
bf	5	0.26%
net	3	0.16%
pub	2	0.11%
info	2	0.11%
gov.bf	2	0.11%
de	2	0.11%

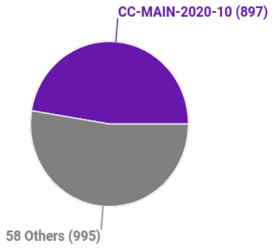
Documents size (in segments) ⓘ

≤ **25** segments **65.96%** (1.2K documents)
> 25 segments **34.04%** (644 documents)



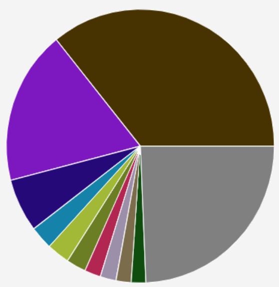
Document collections

CC = 87.37%
IA = 12.63%



Language Distribution

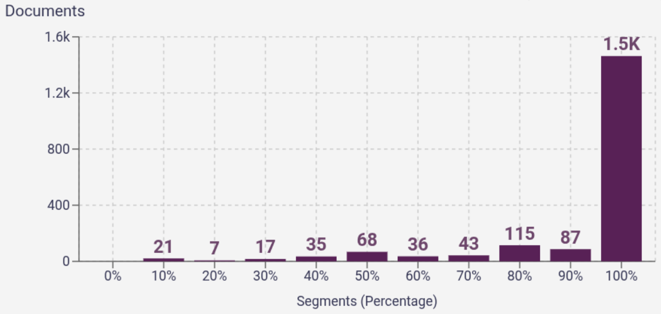
Number of segments in the Mossi corpus



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

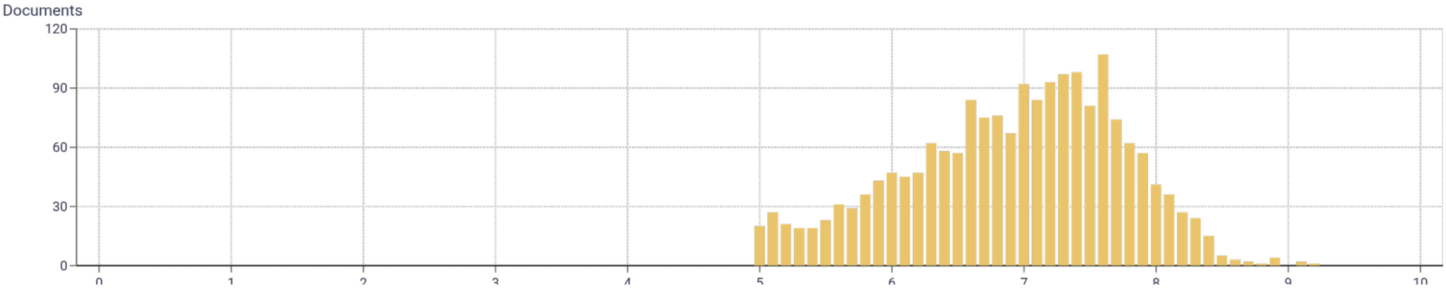
Percentage of segments in Mossi inside documents

segments < 50% - **4.23%** (80 documents)
segments ≥ 50% - **95.77%** (1.8K documents)



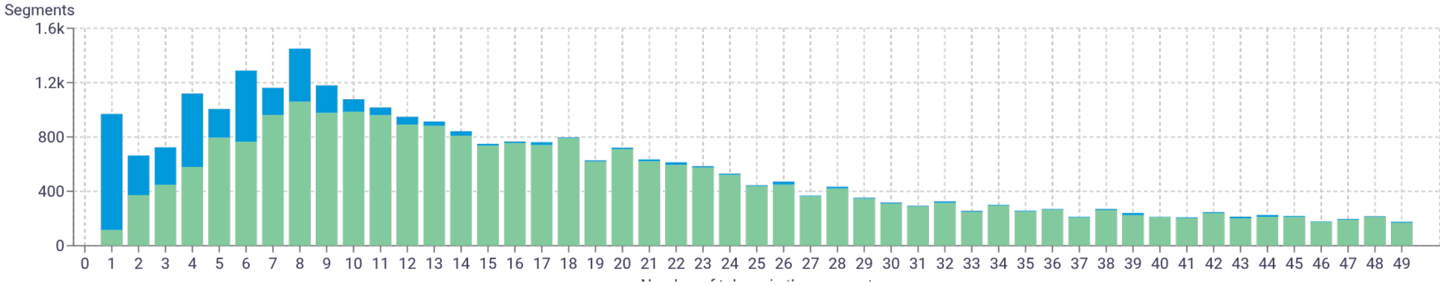
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.9K documents)

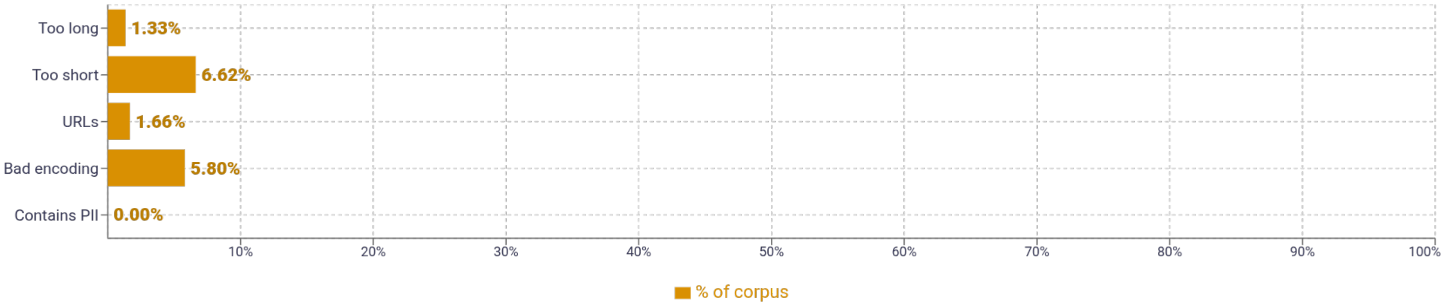


Segment length distribution by token

≤ **49** tokens = **28K** segments | **4.1K** duplicates
> **50** tokens = **20K** segments | **519** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	n | 176,111 sēn | 107,782 tɩ | 71,063 b | 69,763 yaa | 44,636
2	sā n | 15,617 tōe n | 14,945 tɩ b | 14,863 b sēn | 12,529 sēn yaa | 10,981
3	n na n | 5,582 sēn na n | 5,017 d sā n | 3,238 neb nins sēn | 2,663 sēn na yɩl | 2,602
4	sēn na yɩl n | 1,664 b sēn boond tɩ | 979 sēn na yɩl tɩ | 788 bōe yĩng tɩ d | 748 b sēn na n | 740
5	wān to la d tōe | 444 bōe yĩng tɩ d segd | 283 yĩng tɩ d segd n | 282 bōe la d tōe n | 280 būmb ning sēn kɩt tɩ | 247

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av