

General overview

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

Volumes

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

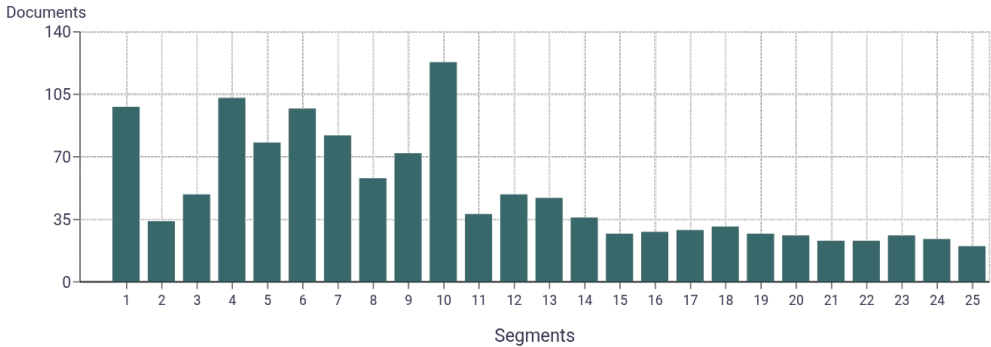
Top 10 domains

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%

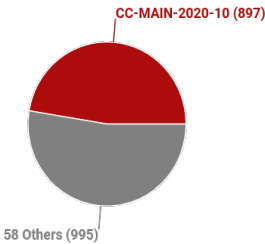
Documents size (in segments) ⓘ



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

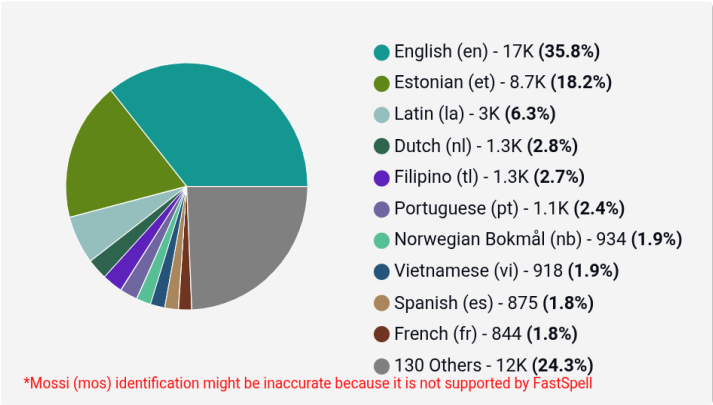
Document collections

CC = **87.37%**  
IA = **12.63%**

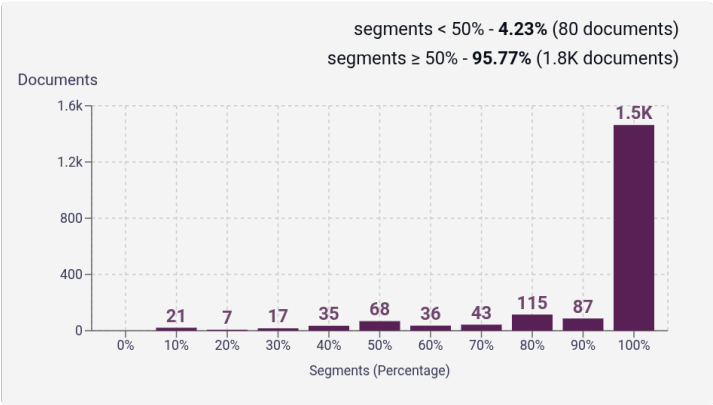


Language Distribution

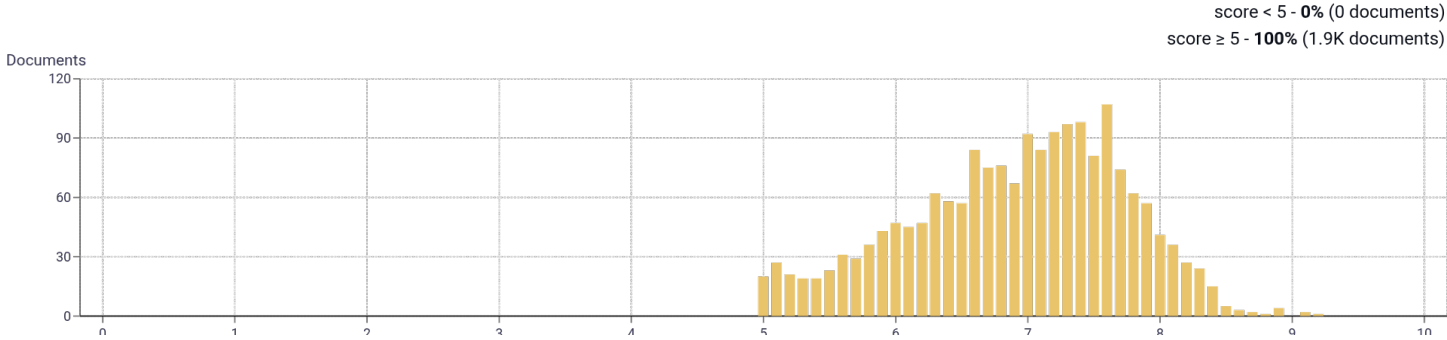
Number of segments in the Mossi (mos) corpus



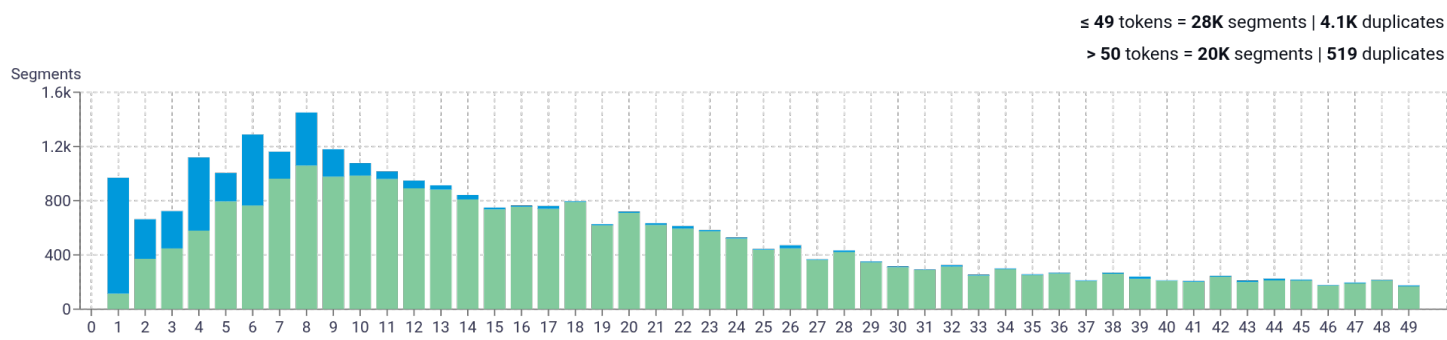
Percentage of segments in Mossi (mos) inside documents



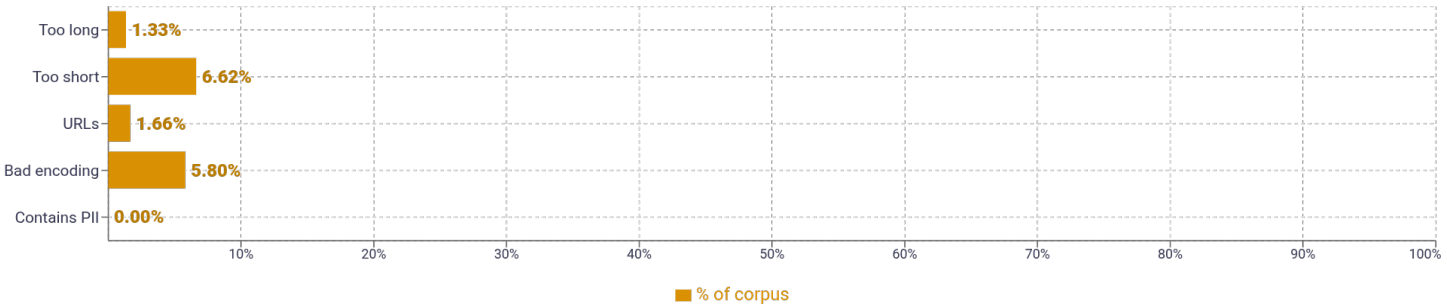
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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>, <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	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				