

General overview

Corpus	Date	Language
hplt-v3-bem_Latn	9/16/2025	Bemba (bem)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5,344	142,918	127,804 (89.42 %)	6.2M	34,074,018	33.43 MB

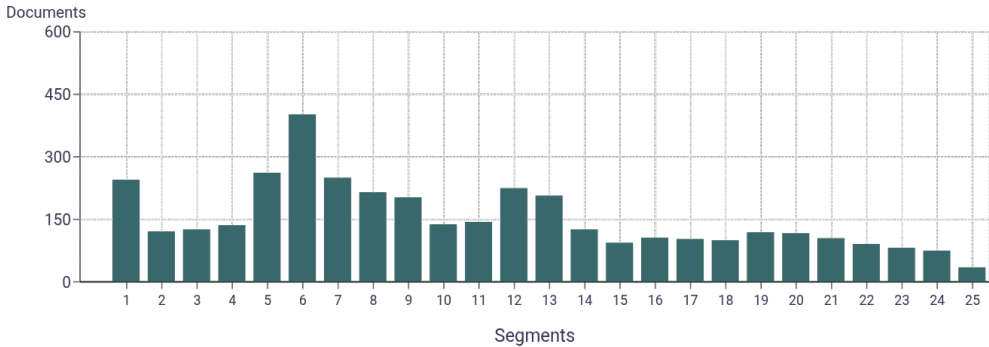
Top 10 domains

Domain	Docs	% of total
jw.org	4.3K	79.60%
worldslastchanc...	406	7.60%
bible.is	231	4.32%
ebible.org	74	1.38%
bible.com	38	0.71%
egw writings.org	28	0.52%
bibleschools.com	26	0.49%
gotquestions.org	23	0.43%
biblearc.com	16	0.30%
33eme-cers.org	16	0.30%

Top 10 TLDs

Domain	Docs	% of total
org	4.5K	83.65%
com	567	10.61%
is	231	4.32%
net	27	0.51%
info	13	0.24%
io	10	0.19%
com.na	5	0.09%
org.za	4	0.07%
co.zw	4	0.07%
bible	3	0.06%

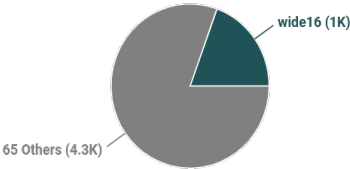
Documents size (in segments) ⓘ



≤ 25 segments 71.61% (3.8K documents)
> 25 segments 28.39% (1.5K documents)

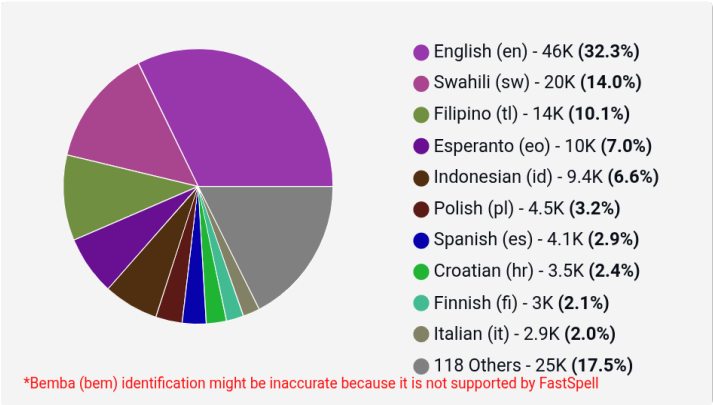
Document collections

CC = 63.10%
IA = 36.90%

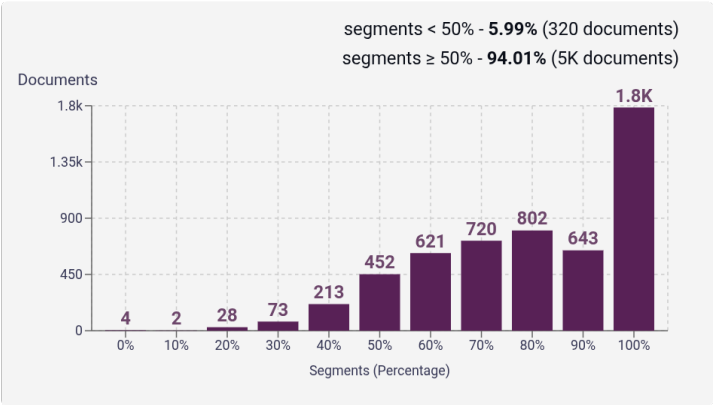


Language Distribution

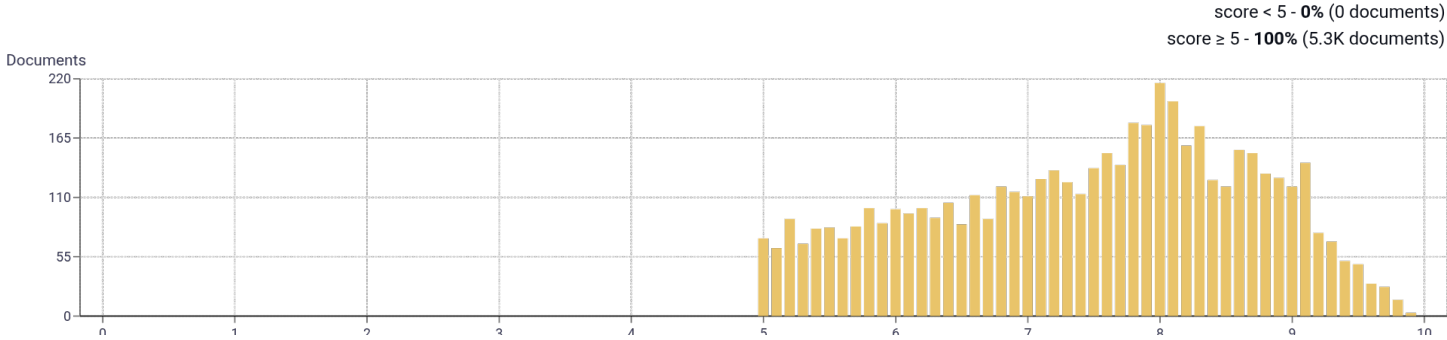
Number of segments in the Bemba (bem) corpus



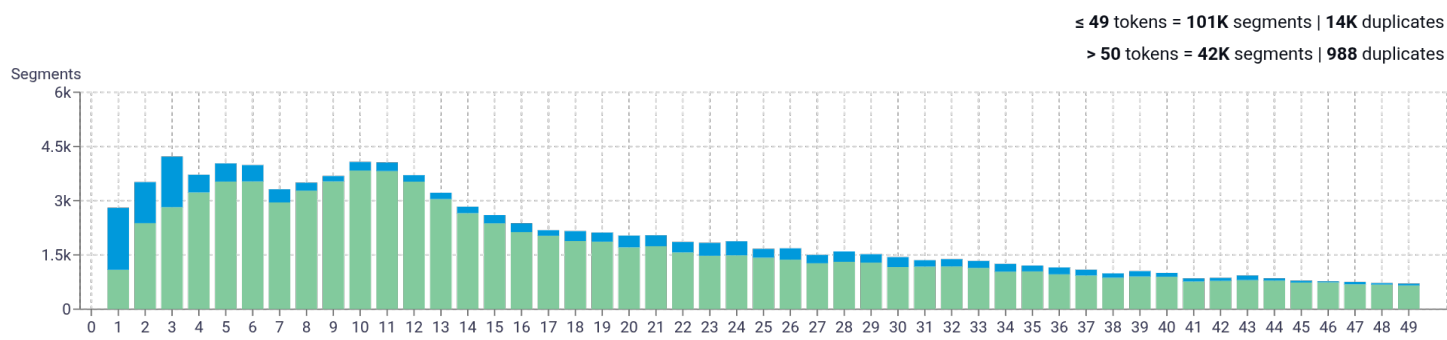
Percentage of segments in Bemba (bem) inside documents



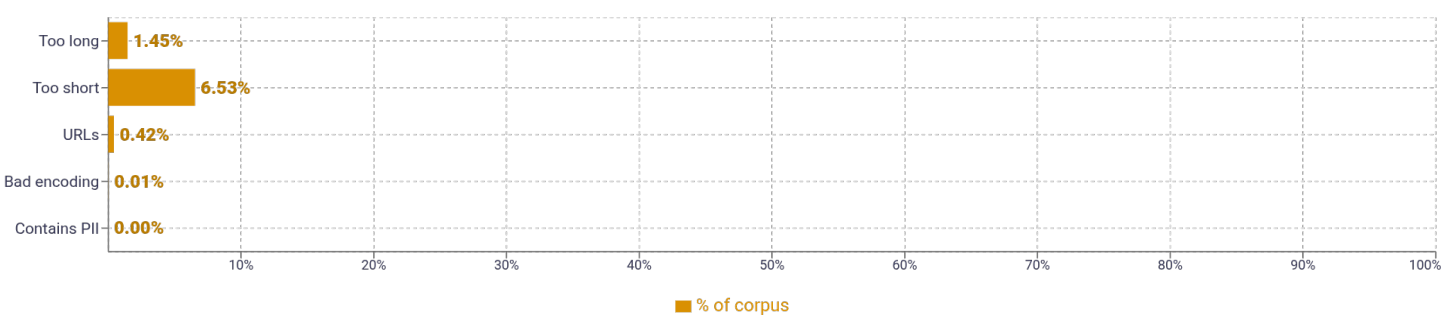
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	a 26,721 yesu 23,426 yehova 15,351 lesa 14,280 mba 14,040	
2	mambo a 1,218 nokuba boobo 1,188 mwami yahuwah 914 i bika 874 calo ca 651	
3	ca kwa lesa 467 batumoni ba yehova 313 i vyani vino 306 cikombelo ca katolika 287 mambo ka o 275	
4	bakamonyi ba kwa yehoba 442 inte sha kwa yehova 235 nte sha kwa yehova 224 ubufumu bwa kwa lesa 213 amalembo ya calo cipya 188	
5	cikombelo ca katolika caku loma 137 akubikka mucibaka cangawo mazina mataanzi 111 mucibaka cangawo mazina mataanzi ngubaapedwe 108 twakagwisya mucibalo citaanzi mazina aabakomba 98 mucibalo citaanzi mazina aabakomba mituni 98	

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				