

General overview

Corpus	Date	Language
hplt-v3-umb_Latn	9/18/2025	Umbundu (umb)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,124	43,464	41,479 (95.43 %)	2.3M	12,003,851	11.56 MB

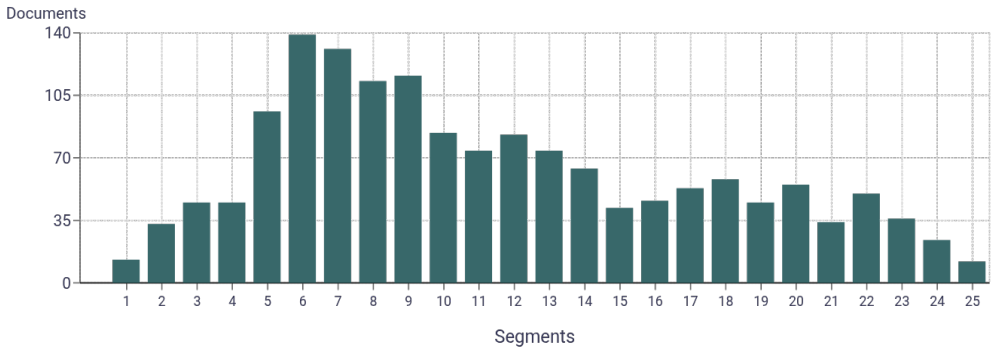
Top 10 domains

Domain	Docs	% of total
jw.org	1.7K	81.78%
kundana.com.na	185	8.71%
neweralive.na	109	5.13%
blogspot.com	19	0.89%
namibian.com.na	14	0.66%
gotquestions.org	10	0.47%
globalrecording...	10	0.47%
omulunga.com.na	7	0.33%
watchtower.org	6	0.28%
stalk.info	2	0.09%

Top 10 TLDs

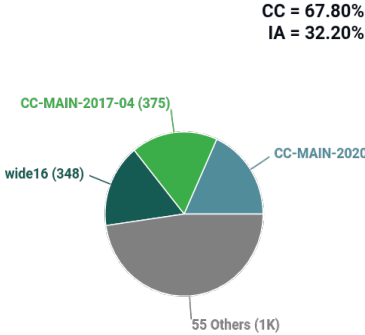
Domain	Docs	% of total
org	1.8K	82.91%
com.na	206	9.70%
na	109	5.13%
com	31	1.46%
net	10	0.47%
info	4	0.19%
ao	2	0.09%
co.uk	1	0.05%

Documents size (in segments) ⓘ



≤ 25 segments **73.68%** (1.6K documents)
> 25 segments **26.32%** (559 documents)

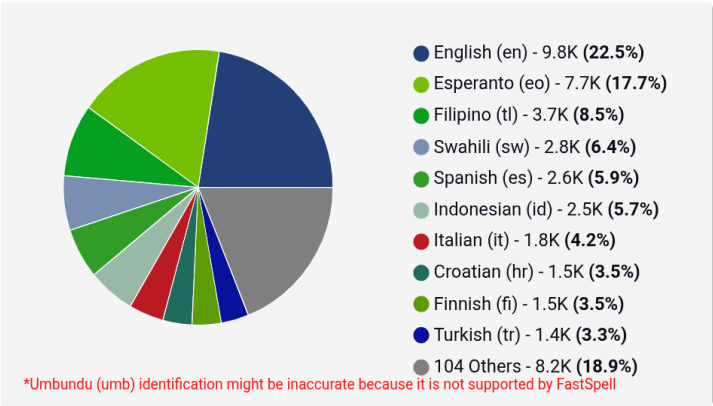
Document collections



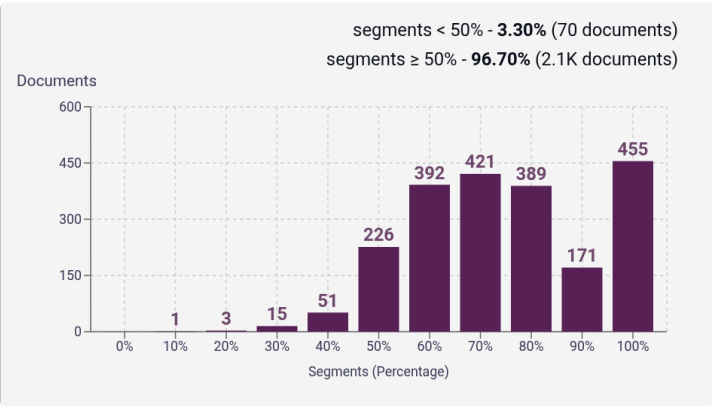
CC = 67.80%
IA = 32.20%

Language Distribution

Number of segments in the Umbundu (umb) corpus

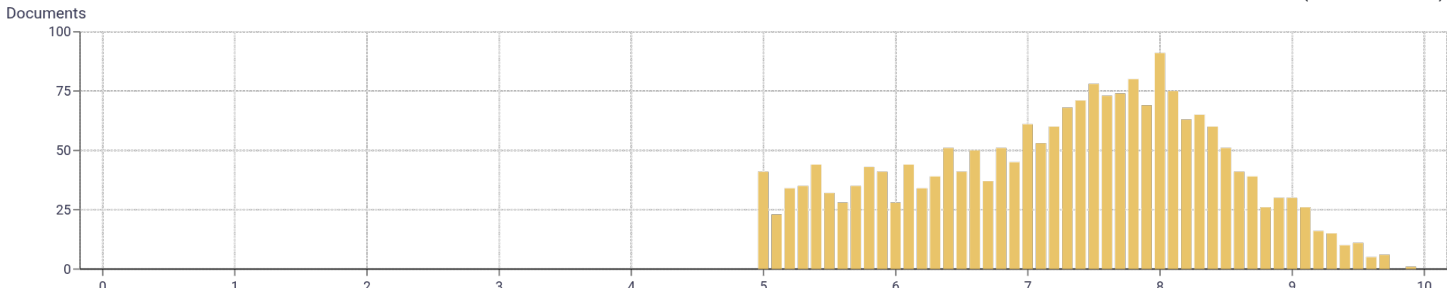


Percentage of segments in Umbundu (umb) inside documents



segments < 50% - **3.30%** (70 documents)
segments ≥ 50% - **96.70%** (2.1K documents)

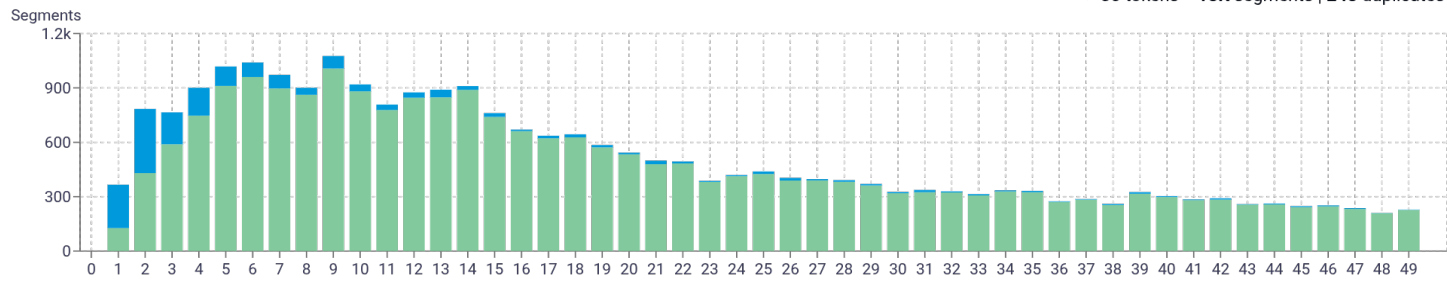
Distribution of documents by document score



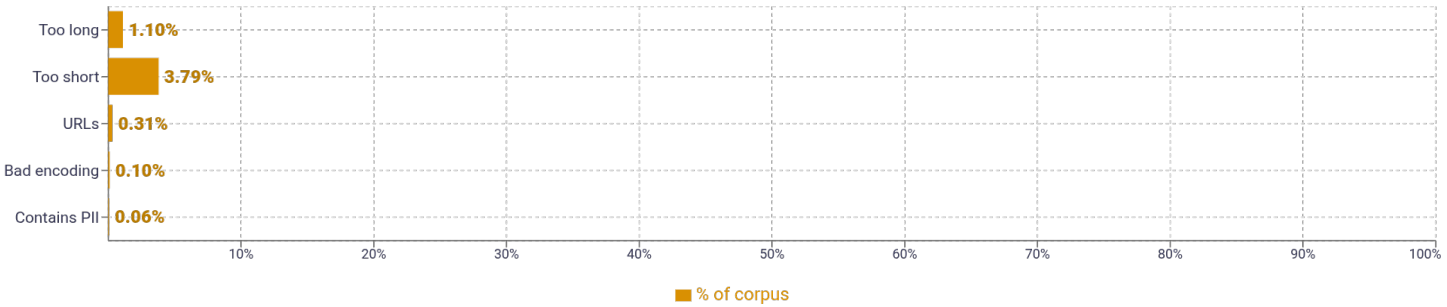
score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (2.1K documents)

Segment length distribution by token

≤ 49 tokens = 26K segments | 1.7K duplicates
> 50 tokens = 18K segments | 248 duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	li 47,018 va 37,946 ku 35,244 ka 23,735 tu 19,348	
2	okwa li 9,835 ova li 4,760 li va 3,979 nosho yo 2,829 va li 2,603	
3	ova li va 2,570 shi na sha 1,440 jesus okwa li 1,367 ovo va li 963 va li va 856	
4	okwa li a lombwela 401 okwa li e va 339 ovo va li va 276 va kele na ku 267 okwa li a hala 249	
5	kashi na nee mbudi kutya 217 dulu oku tu kwafela tu 167 mbela ou shi shii kutya 140 okwa li e shii kutya 134 jesus okwa li a lombwela 102	

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				