

General overview

Corpus	Date	Language
hplt-v3-nya_Latn	9/18/2025	Nyanja (ny)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
177,891	4,292,677	3,357,457 (78.21 %)	106M	657,291,550	630.75 MB

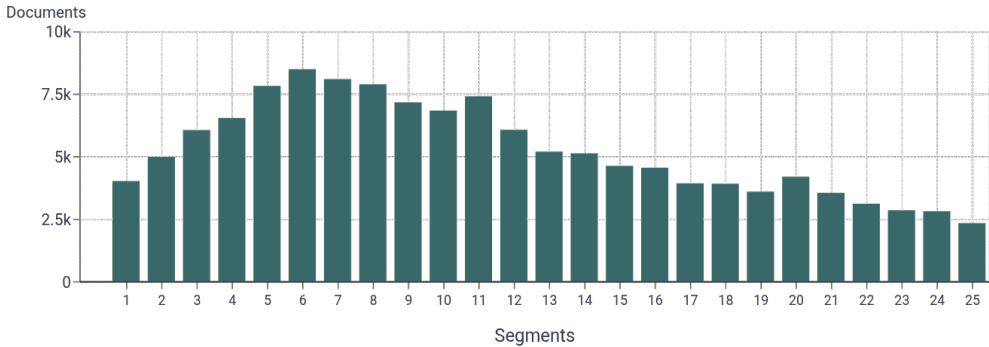
Top 10 domains

Domain	Docs	% of total
reviews.tn	8.1K	4.58%
jw.org	6.4K	3.60%
eturbonews.com	5.3K	2.98%
wondershare.com	5.1K	2.88%
martech.zone	4.9K	2.75%
androidsis.com	3.7K	2.06%
manuals.plus	3.4K	1.92%
actualidadiphon...	3.2K	1.78%
actualidadgadg...	2.8K	1.56%
recetin.com	2.6K	1.47%

Top 10 TLDs

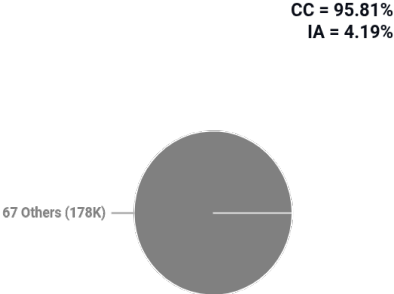
Domain	Docs	% of total
com	133K	74.94%
org	11K	6.33%
tn	8.1K	4.58%
zone	4.9K	2.75%
plus	3.4K	1.92%
net	3.2K	1.81%
ru	1.3K	0.73%
mw	1.1K	0.63%
news	681	0.38%
es	615	0.35%

Documents size (in segments) ⓘ



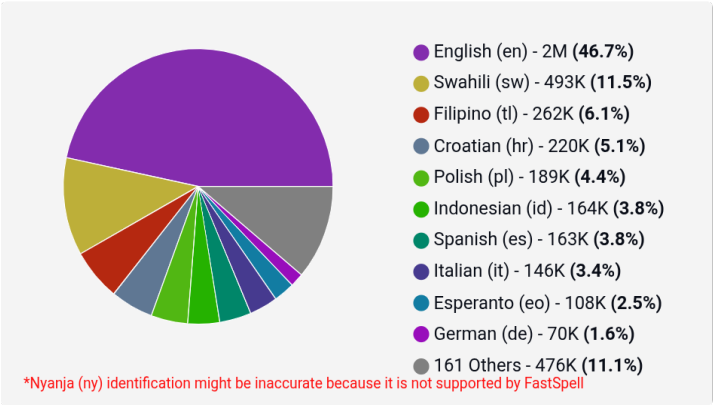
≤ 25 segments 73.9% (131K documents)
> 25 segments 26.1% (46K documents)

Document collections

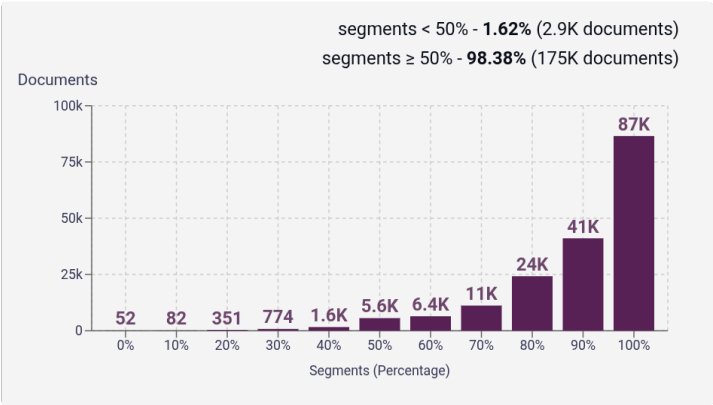


Language Distribution

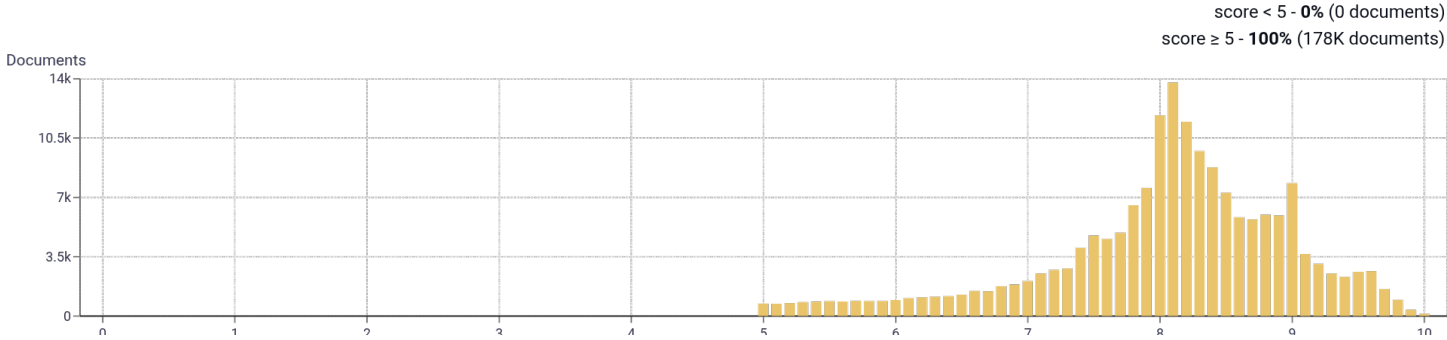
Number of segments in the Nyanja (ny) corpus



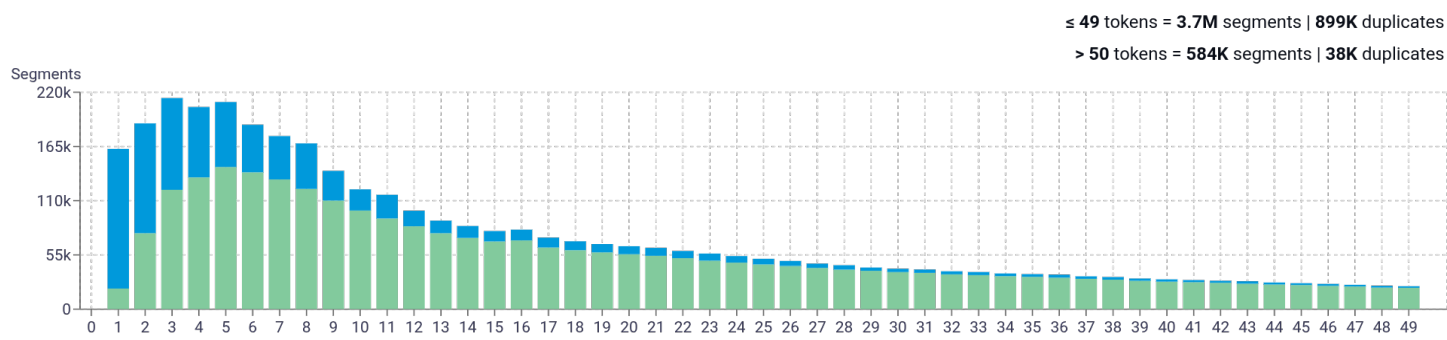
Percentage of segments in Nyanja (ny) inside documents



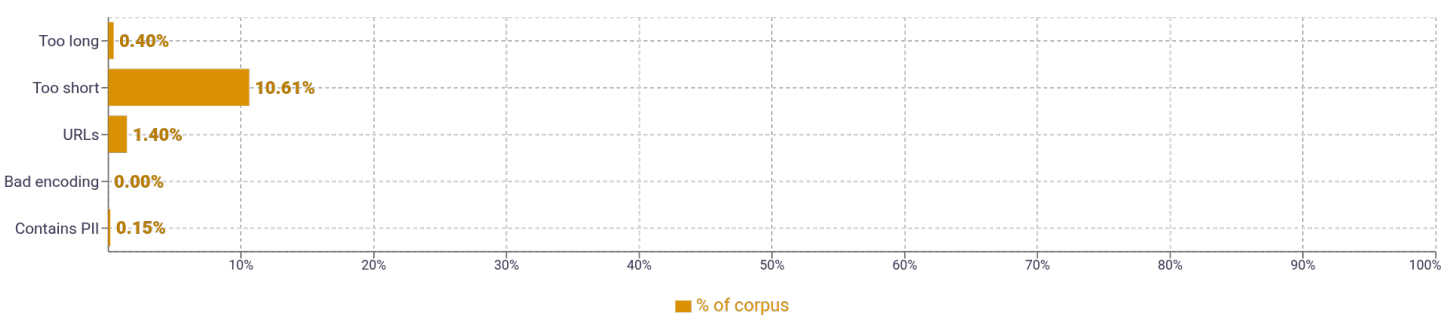
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ya 609,324 m 606,216 kwambiri 579,638 ndipo 577,162 komanso 571,415	
2	nthawi zambiri 62,770 nthawi zonse 59,912 lonse lapansi 57,606 omwe ali 50,288 padziko lonse 45,393	
3	padziko lonse lapansi 37,292 khalani oyamba kuyankha 12,875 kayendetsedwe ka kayendetsedwe 10,913 tsiku ndi tsiku 10,563 ka kayendetsedwe ka 10,424	
4	kayendetsedwe ka kayendetsedwe ka 10,344 ka kayendetsedwe ka kayendetsedwe 10,150 kugawana nawo nkhani yathu 7,308 pamasamba ochezera kuti atilimbikitse 7,307 nkhani yathu pamasamba ochezera 7,307	
5	kayendetsedwe ka kayendetsedwe ka kayendetsedwe 10,097 ka kayendetsedwe ka kayendetsedwe ka 9,757 yathu pamasamba ochezera kuti atilimbikitse 7,307 nawo nkhani yathu pamasamba ochezera 7,307 kugawana nawo nkhani yathu pamasamba 7,307	

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				