HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-tum_Latn	9/18/2025	Tumbuka

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5,654	157,200	143,819 (91.49 %)	5.8M	32,873,782	32.07 MB

Top 10 domains

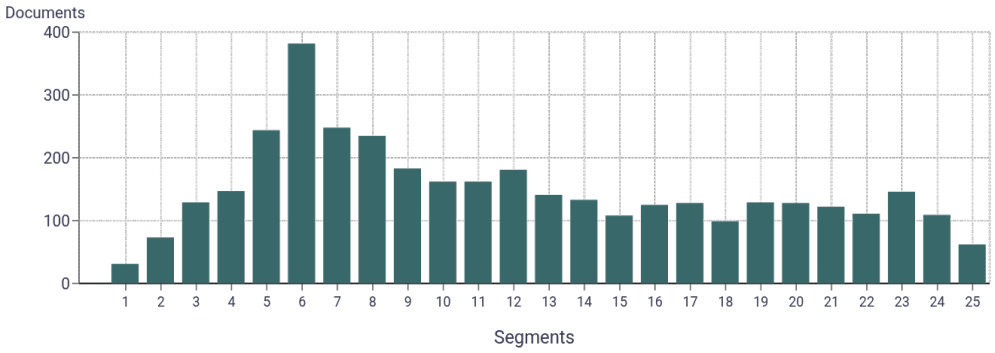
Domain	Docs	% of total
jw.org	5K	88.77%
wikipedia.org	370	6.54%
nkhanimchitumbu...	58	1.03%
wordpress.com	33	0.58%
islamhouse.com	18	0.32%
fountainofvicto...	17	0.30%
globalrecording...	15	0.27%
bywiki.com	13	0.23%
quranenc.com	11	0.19%
bible.com	11	0.19%

Top 10 TLDs

Domain	Docs	% of total
org	5.4K	96.27%
com	172	3.04%
net	18	0.32%
is	5	0.09%
mw	2	0.04%
io	2	0.04%
co.uk	2	0.04%
wiki	1	0.02%
vn	1	0.02%
tech	1	0.02%

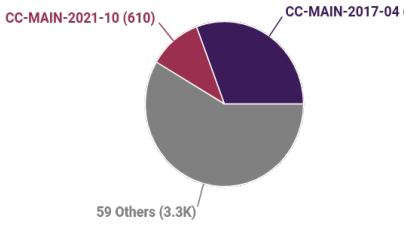
Documents size (in segments)

≤ 25 segments **65.76%** (3.7K documents)
> 25 segments **34.24%** (1.9K documents)



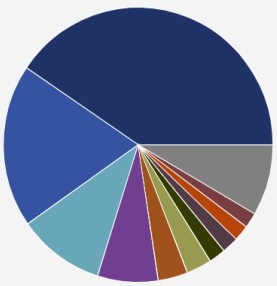
Document collections

CC = 93.37%
IA = 6.63%



Language Distribution

Number of segments in the Tumbuka corpus

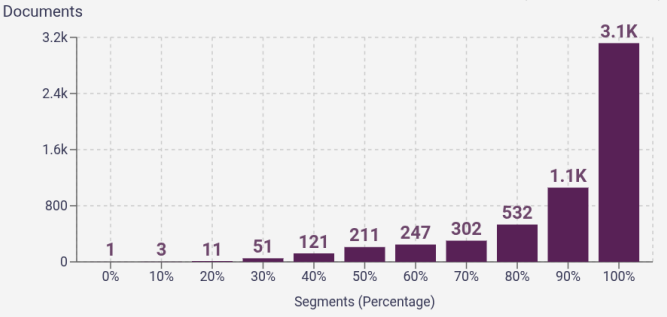


*Tumbuka identification might be inaccurate because it is not supported by FastSpell.

- English - 64K **(40.5%)**
- Swahili - 30K **(19.2%)**
- Filipino - 16K **(10.4%)**
- Indonesian - 11K **(7.3%)**
- Spanish - 5.6K **(3.6%)**
- Italian - 4.7K **(3.0%)**
- Esperanto - 3.2K **(2.0%)**
- Waray - 3.1K **(2.0%)**
- Polish - 2.9K **(1.9%)**
- Croatian - 2.8K **(1.8%)**
- 116 Others - 13K **(8.4%)**

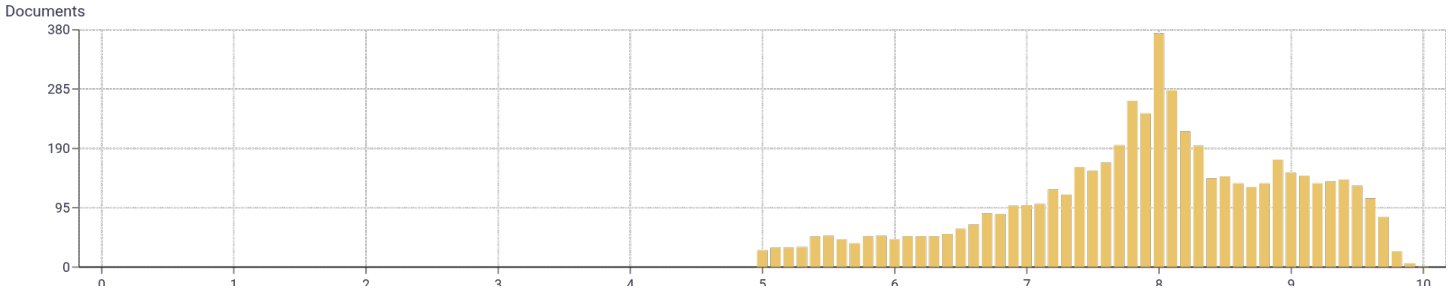
Percentage of segments in Tumbuka inside documents

segments < 50% - **3.31%** (187 documents)
segments ≥ 50% - **96.69%** (5.5K documents)



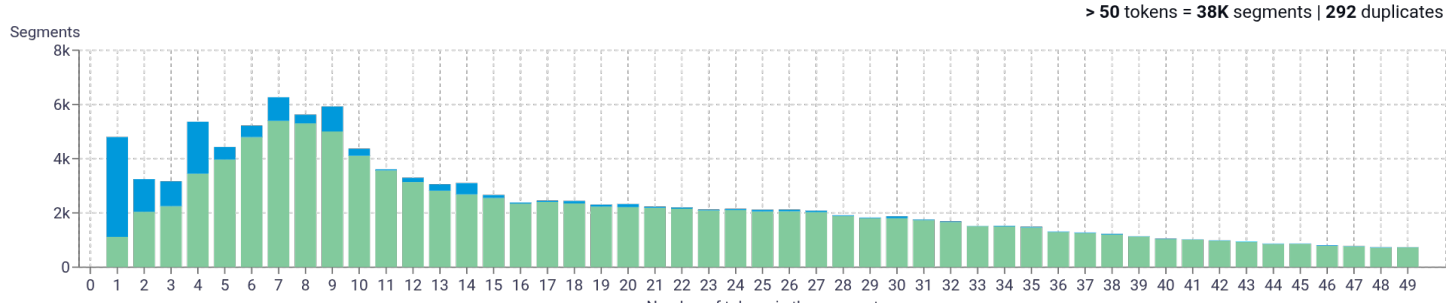
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (5.7K documents)

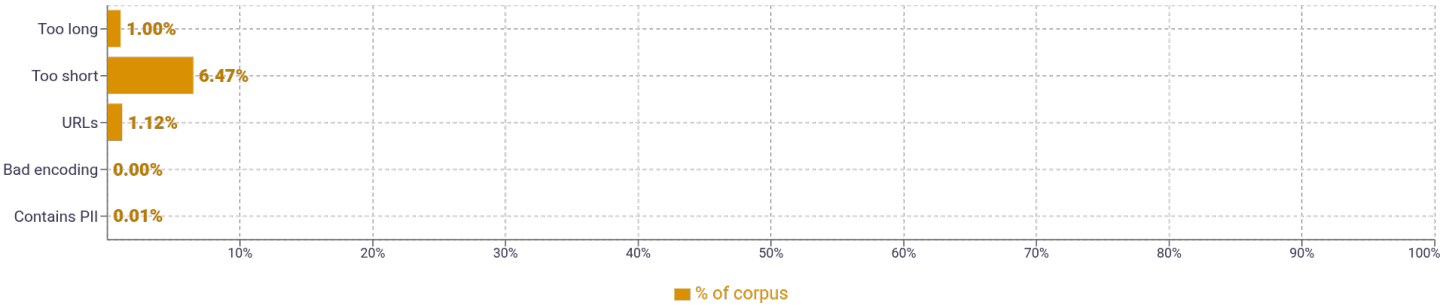


Segment length distribution by token

≤ 49 tokens = **119K** segments | **13K** duplicates
> 50 tokens = **38K** segments | **292** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ŵanthu | 46,274 · ndi | 38,099 · yehova | 33,372 · chiuta | 25,995 · yesu | 22,127
2	ŵanthu ŵa | 4,920 · ŵanthu ŵanandi | 3,267 · ŵanthu wo | 2,535 · from the | 2,356 · the original | 2,201
3	from the original | 2,195 · archived from the | 2,066 · the original on | 2,035 · ŵakaboni ŵa yehova | 1,504 · anamuliro gha charu | 1,159
4	archived from the original | 2,066 · from the original on | 2,035 · anamuliro gha charu chiphya | 1,159 · malemba ghakupatulika mu mang | 1,045 · baibolo la pa intaneti | 1,016
5	archived from the original on | 1,906 · kumbi ndi vinthu wuli vo | 243 · vyaru vya kumanjiliro gha dazi | 137 · sambiru la bayibolu la mpingu | 108 · ndicho fumu yikuru yehova yayowoya | 98

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
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				