HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-nya_Latn	9/18/2025	Chichewa

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

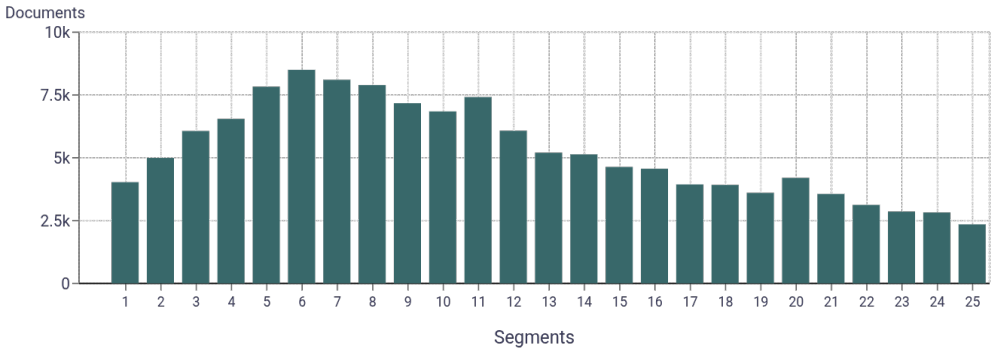
Domain	Docs	% of total
reviews.tn	8.1K	4.58%
jw.org	6.4K	3.60%
eturbonews.com	5.3K	2.98%
wondershare.com	5.1K	2.88%
martech.zone	4.9K	2.75%
androidsis.com	3.7K	2.06%
manuals.plus	3.4K	1.92%
actualidadiphon...	3.2K	1.78%
actualidadgadge...	2.8K	1.56%
recetin.com	2.6K	1.47%

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%

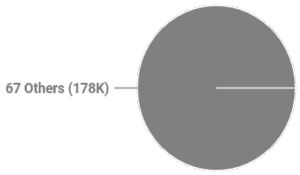
Documents size (in segments)

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



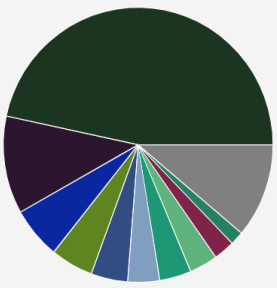
Document collections

CC = 95.81%
IA = 4.19%



Language Distribution

Number of segments in the Chichewa corpus

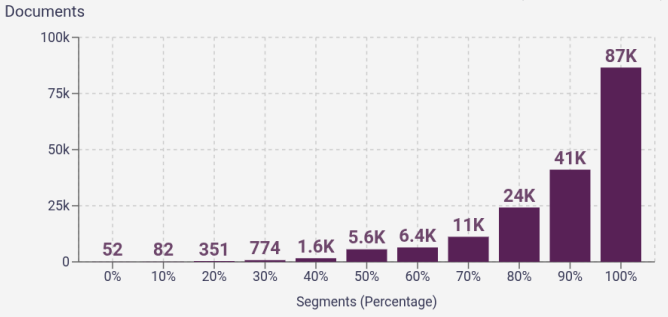


*Chichewa identification might be inaccurate because it is not supported by FastSpell.

- English - 2M **(46.7%)**
- Swahili - 493K **(11.5%)**
- Filipino - 262K **(6.1%)**
- Croatian - 220K **(5.1%)**
- Polish - 189K **(4.4%)**
- Indonesian - 164K **(3.8%)**
- Spanish - 163K **(3.8%)**
- Italian - 146K **(3.4%)**
- Esperanto - 108K **(2.5%)**
- German - 70K **(1.6%)**
- 161 Others - 476K **(11.1%)**

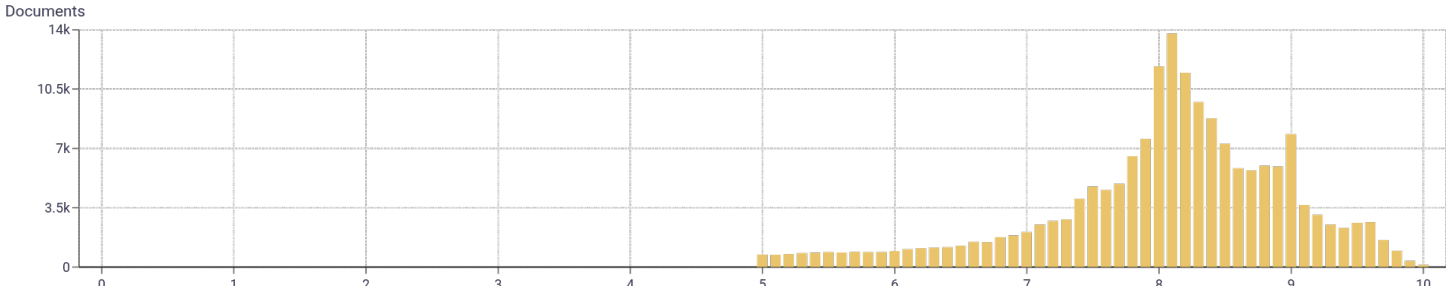
Percentage of segments in Chichewa inside documents

segments < 50% - **1.62%** (2.9K documents)
segments ≥ 50% - **98.38%** (175K documents)



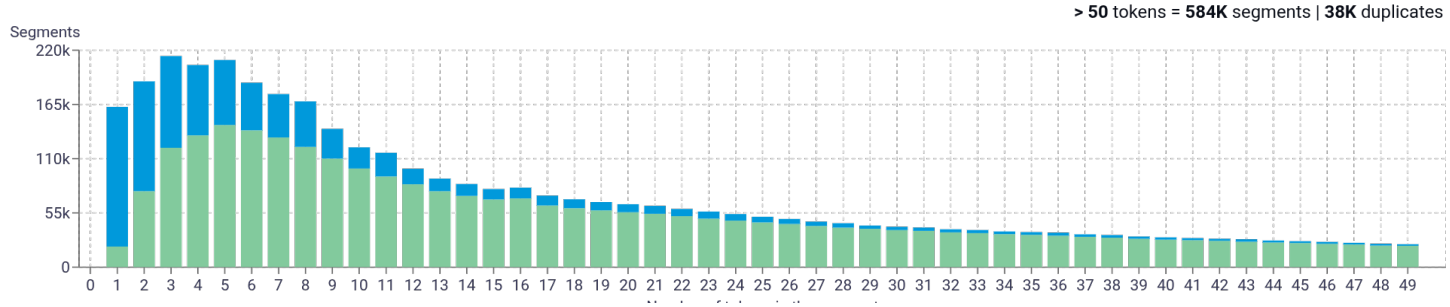
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (178K documents)

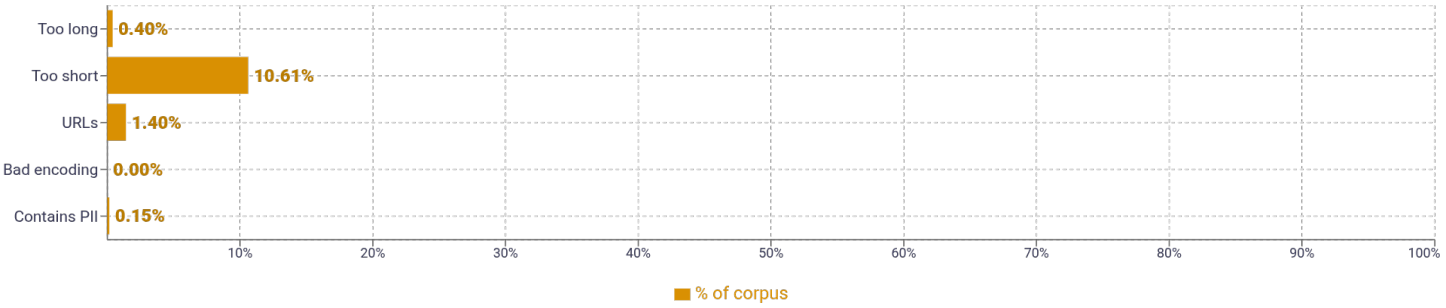


Segment length distribution by token

≤ 49 tokens = **3.7M** segments | **899K** duplicates
> 50 tokens = **584K** segments | **38K** duplicates



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>, <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				