HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-cjk_Latn	9/17/2025	Chokwe (cjk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,081	29,646	27,503 (92.77 %)	1.2M	6,972,103	6.73 MB

Top 10 domains

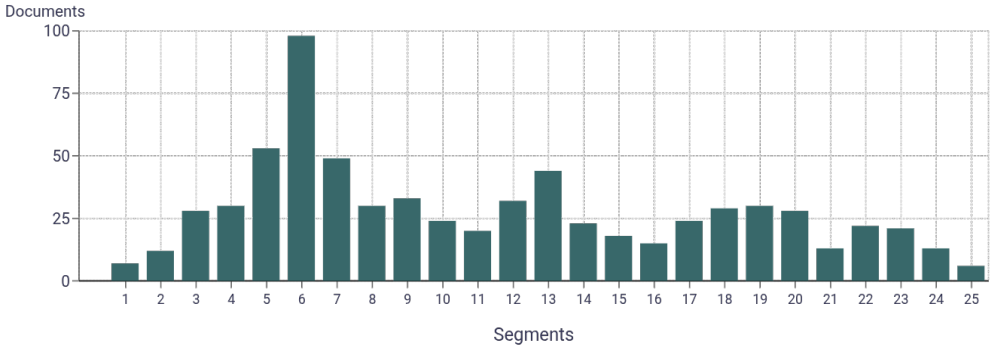
Domain	Docs	% of total
jw.org	1K	96.39%
globalrecording...	9	0.83%
bible.com	7	0.65%
watchtower.org	4	0.37%
rna.ao	4	0.37%
contafrica.org	3	0.28%
unicode.org	2	0.19%
svfellowship.info	1	0.09%
securesites.net	1	0.09%
ohchr.org	1	0.09%

Top 10 TLDs

Domain	Docs	% of total
org	1.1K	97.59%
net	10	0.93%
com	10	0.93%
ao	4	0.37%
info	1	0.09%
gq	1	0.09%

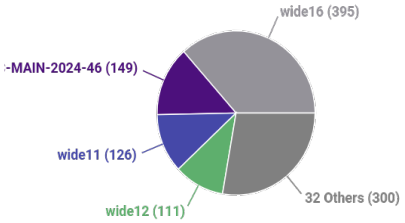
Documents size (in segments)

≤ 25 segments **64.94%** (702 documents)
> 25 segments **35.06%** (379 documents)



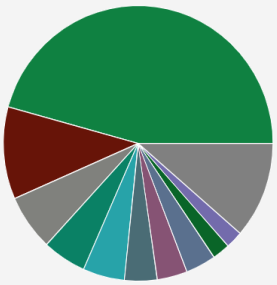
Document collections

CC = 37.37%
IA = 62.63%



Language Distribution

Number of segments in the Chokwe (cjk) corpus

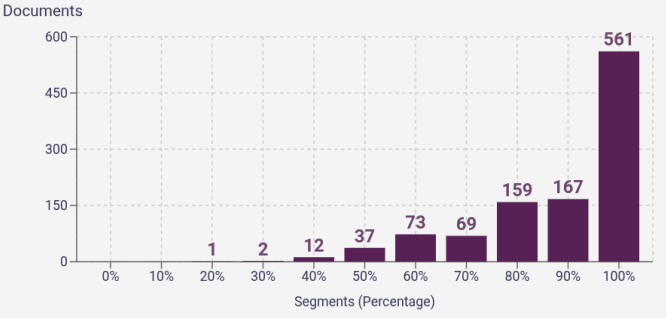


- English (en) - 14K **(45.7%)**
- Swahili (sw) - 3.2K **(10.8%)**
- Filipino (tl) - 1.9K **(6.6%)**
- Indonesian (id) - 1.6K **(5.2%)**
- Spanish (es) - 1.5K **(5.0%)**
- Italian (it) - 1.2K **(3.9%)**
- Finnish (fi) - 1.1K **(3.6%)**
- Esperanto (eo) - 1.1K **(3.6%)**
- German (de) - 623 **(2.1%)**
- Polish (pl) - 608 **(2.1%)**
- 83 Others - 3.4K **(11.3%)**

*Chokwe (cjk) identification might be inaccurate because it is not supported by FastSpell

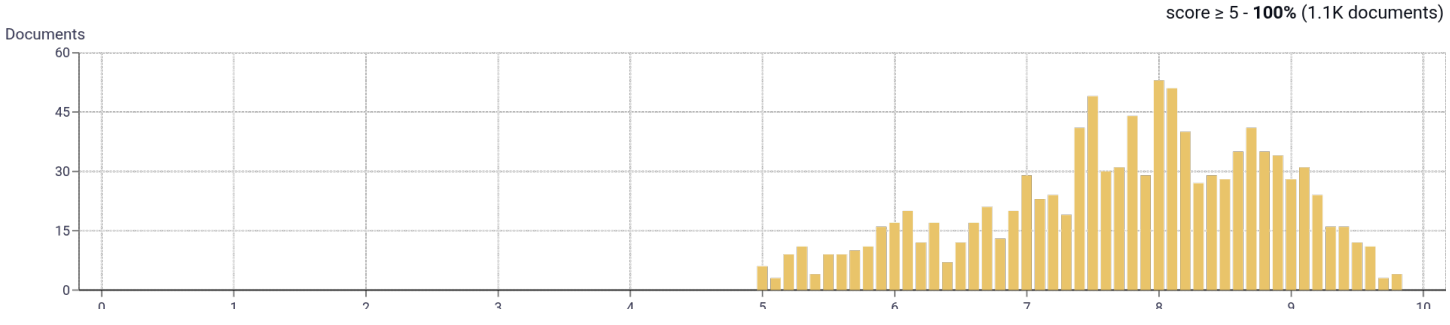
Percentage of segments in Chokwe (cjk) inside documents

segments < 50% - **1.39%** (15 documents)
segments ≥ 50% - **98.61%** (1.1K documents)



Distribution of documents by document score

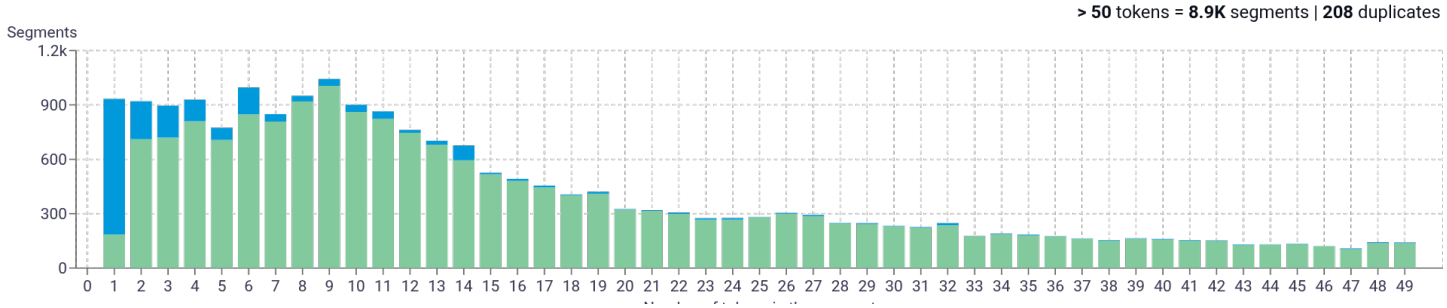
score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.1K documents)



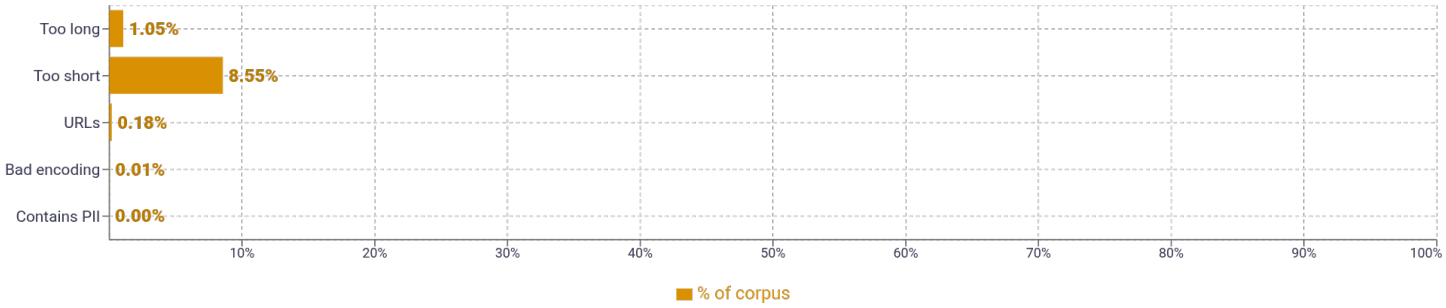
Segment length distribution by token

≤ 49 tokens = **21K** segments | **1.9K** duplicates

> 50 tokens = **8.9K** segments | **208** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	yehova | 9,972 · mu | 8,275 · ha | 7,330 · nawa | 7,248 · wa | 6,777
2	mumu liaka | 833 · vyuma muka | 795 · mu mbimbiliya | 680 · hakutwala ku | 670 · kaha nawa | 666
3	yela ja yehova | 374 · wanangana wa zambi | 281 · liji lia zambi | 228 · jina lia zambi | 181 · haya myaka yosena | 177
4	mu zuwo lia ususu | 132 · ku miaka ya mutolo | 80 · mu mulimo wa kwambujola | 73 · paraisu ya ku spiritu | 72 · mu liji lia zambi | 72
5	world translation of the holy | 70 · translation of the holy scriptures | 70 · new world translation of the | 66 · yisoneko ya ngregu ya akwa | 49 · mutuhasa kupwa ni shindakenyo ngwetu | 49

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				