HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-twi_Latn	9/18/2025	Twi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7,896	213,527	177,799 (83.27 %)	8.4M	35,414,855	38 MB

Top 10 domains

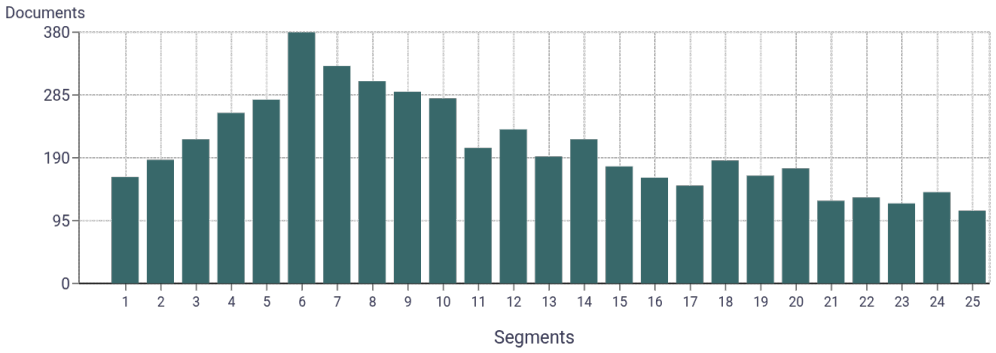
Domain	Docs	% of total
jw.org	2.6K	33.41%
biblica.com	1.2K	15.08%
wikipedia.org	911	11.54%
ebible.org	560	7.09%
biblegateway.com	406	5.14%
barbersupplyand...	223	2.82%
bible.is	102	1.29%
4jehovah.org	101	1.28%
amoafowaa.com	96	1.22%
cupkin.com	89	1.13%

Top 10 TLDs

Domain	Docs	% of total
org	4.7K	59.19%
com	2.8K	35.40%
is	103	1.30%
net	88	1.11%
ru	59	0.75%
store	32	0.41%
com.au	13	0.16%
cc	11	0.14%
cat	11	0.14%
org.uk	10	0.13%

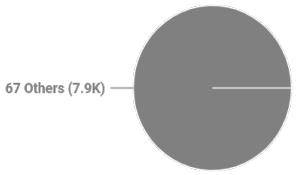
Documents size (in segments)

≤ 25 segments **65.44%** (5.2K documents)
> 25 segments **34.56%** (2.7K documents)



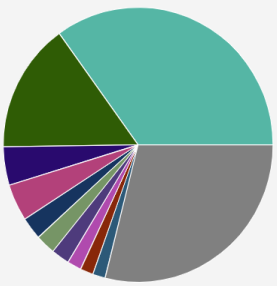
Document collections

CC = 92.88%
IA = 7.12%



Language Distribution

Number of segments in the Twi corpus

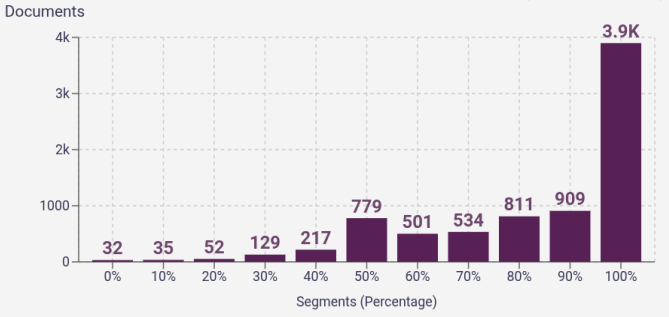


*Twi identification might be inaccurate because it is not supported by FastSpell.

- English - 75K **(35.0%)**
- Spanish - 33K **(15.3%)**
- Esperanto - 9.7K **(4.5%)**
- Polish - 9.3K **(4.4%)**
- Italian - 5.7K **(2.6%)**
- Finnish - 5.1K **(2.4%)**
- Amharic - 4.6K **(2.2%)**
- Hungarian - 3.6K **(1.7%)**
- Slovenian - 3.4K **(1.6%)**
- French - 3.2K **(1.5%)**
- 154 Others - 62K **(28.9%)**

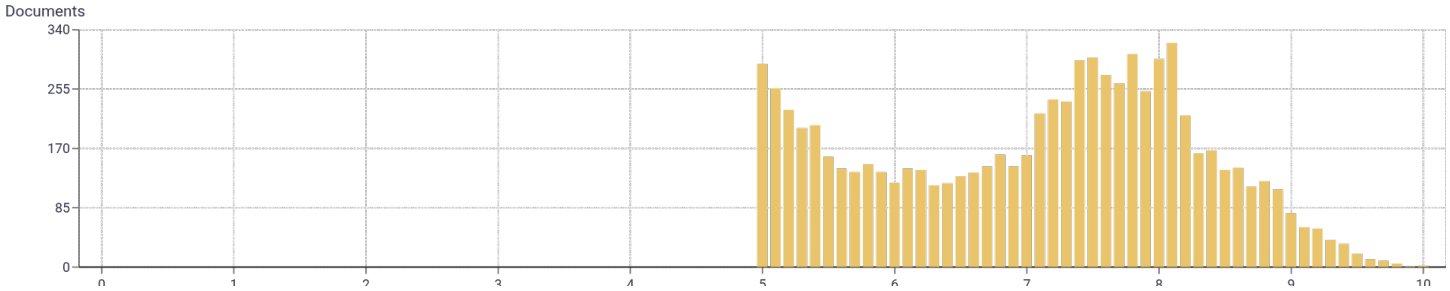
Percentage of segments in Twi inside documents

segments < 50% - **5.89%** (465 documents)
segments ≥ 50% - **94.11%** (7.4K documents)



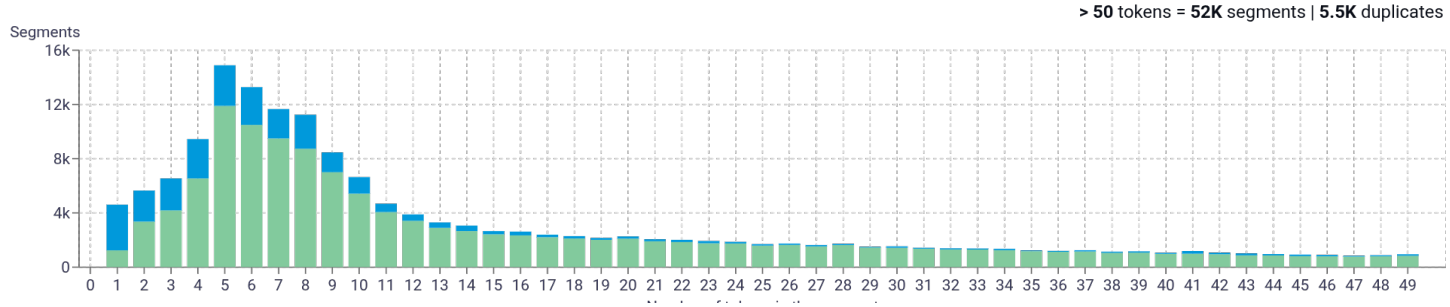
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (7.9K documents)

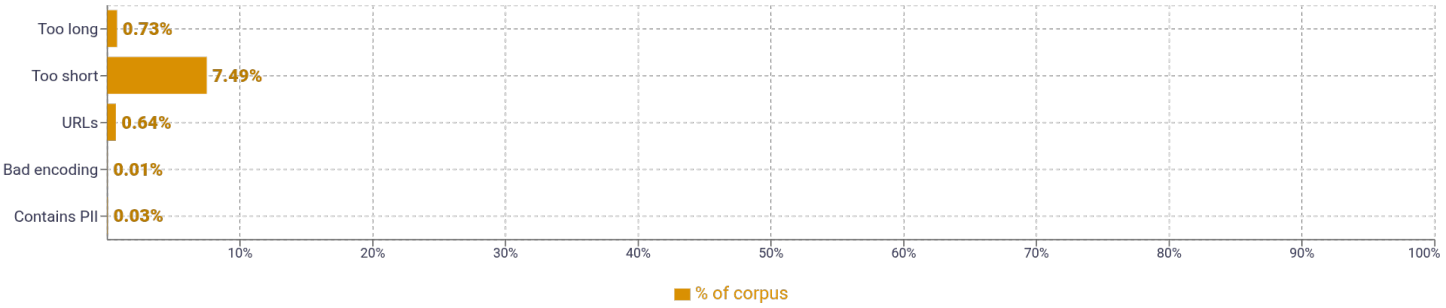


Segment length distribution by token

≤ 49 tokens = **162K** segments | **30K** duplicates
> 50 tokens = **52K** segments | **5.5K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	mu | 112,855 · wɔ | 104,799 · wɔn | 65,722 · wo | 41,486 · ma | 36,820
2	in the | 5,133 · yi mu | 3,414 · di dwuma | 3,213 · ɔwɔ kɛ | 3,129 · mu wɔ | 2,973
3	wɔde di dwuma | 1,155 · wɔ mekɛ mɔɔ | 939 · dodow no ara | 870 · saa asɛm yi | 745 · nneɛma a wɔde | 677
4	atwa yɛn ho ahyia | 352 · archived from the original | 317 · wɔ afe apem ahankron | 304 · wei ɛwɔ asante kasa | 298 · from the original on | 290
5	wɔatwerɛ nsɛm wei ɛwɔ asante | 298 · wei ɛwɔ asante kasa mu | 298 · archived from the original on | 280 · duzu ati a ɔwɔ kɛ | 173 · yi wɔ akuapem kasa mu | 166

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				