

General overview

Corpus	Date	Language
hplt-v3-twi_Latn	9/18/2025	Akan (tw)

Volumes

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

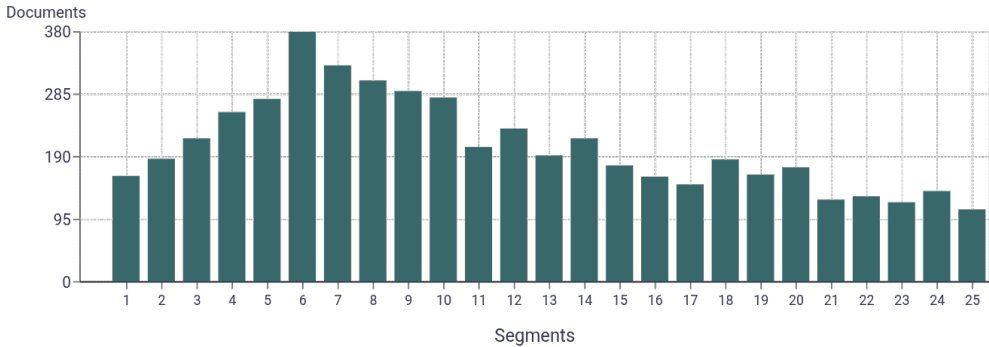
Top 10 domains

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%

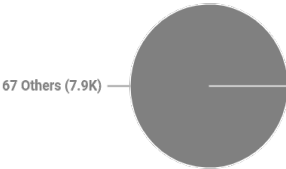
Documents size (in segments) ⓘ



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

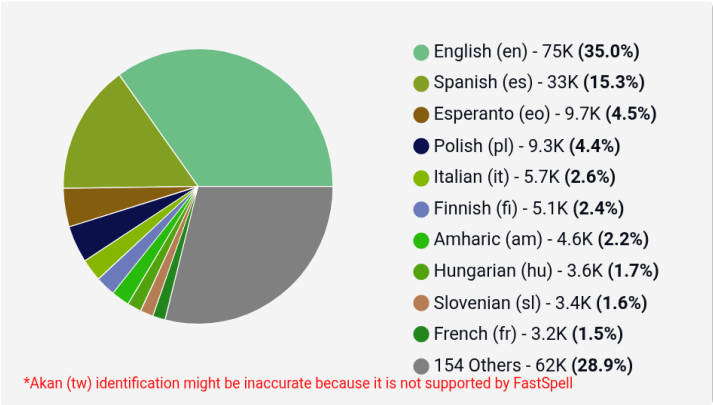
Document collections

CC = 92.88%
IA = 7.12%

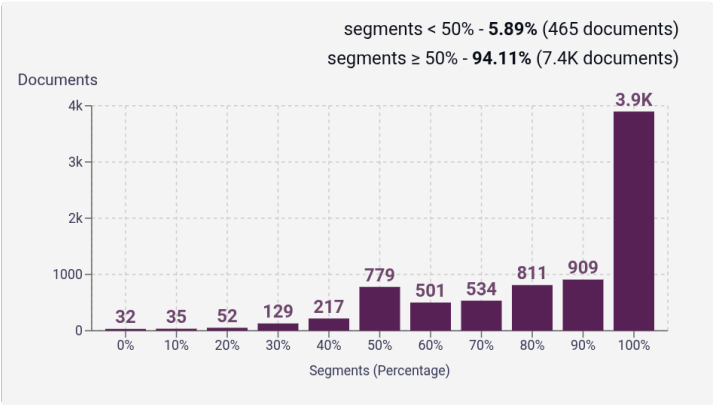


Language Distribution

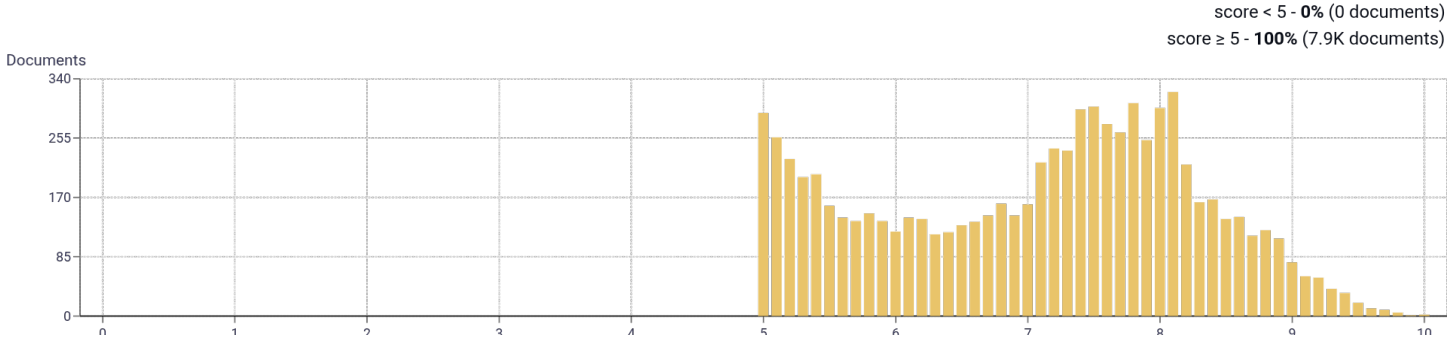
Number of segments in the Akan (tw) corpus



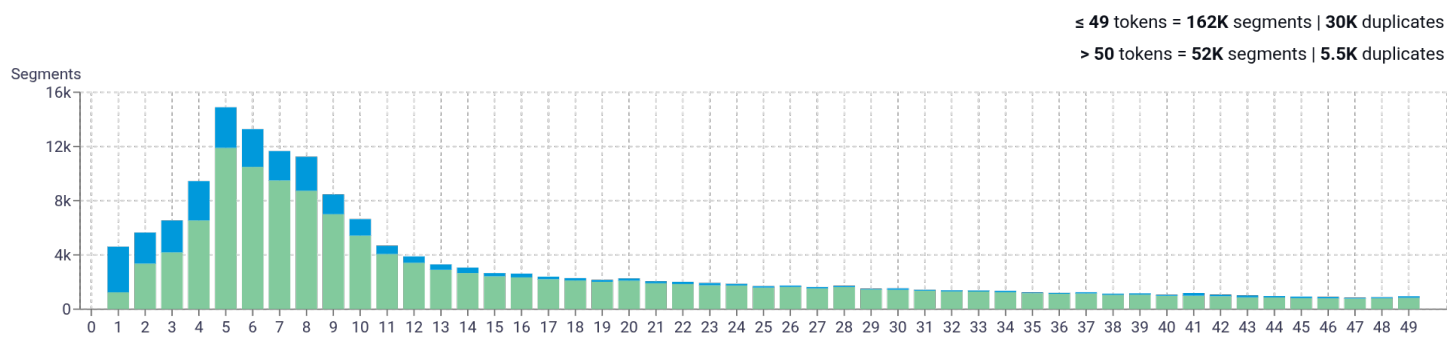
Percentage of segments in Akan (tw) inside documents



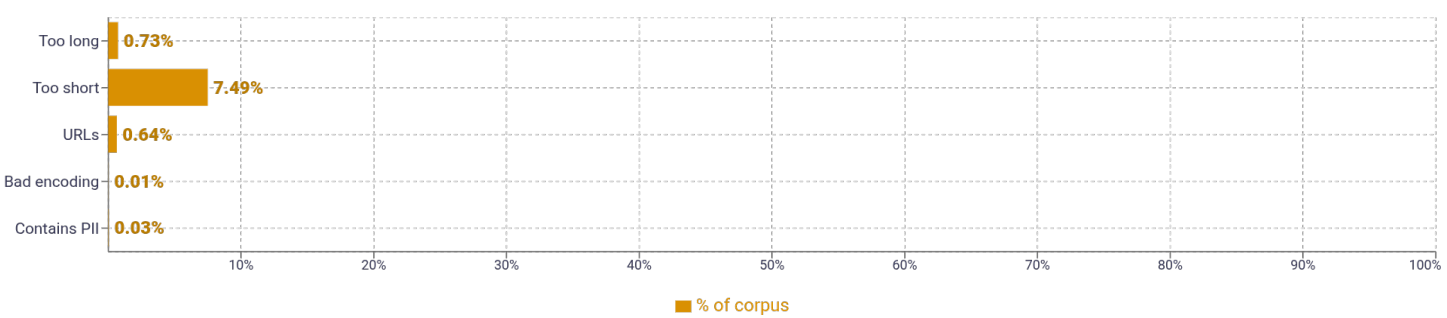
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	mu 112,855 כּוּ 104,799 חכּוּ 65,722 וּ 41,486 מַ 36,820	
2	in the 5,133 יי mu 3,414 di dwuma 3,213 כּוּ כּוּ 3,129 mu כּוּ 2,973	
3	wode di dwuma 1,155 כּוּ meכּוּ 939 dodow no ara 870 saa asem yi 745 nneema a wode 677	
4	atwa yen ho ahya 352 archived from the original 317 כּוּ afe apem ahankron 304 wei ewo asante kasa 298 from the original on 290	
5	watwerε nsem wei ewo asante 298 wei ewo asante kasa mu 298 archived from the original on 280 duzu ati a כּוּ כּוּ 173 yi כּוּ akuapem kasa mu 166	

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				