

General overview

Corpus	Analytics date	Language
twi_Latn.jsonl.tsv	12/5/2024	Akan (tw)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
5,860	125,611	71,959 (57.29 %)	5.7M	25.97 MB	24,057,642

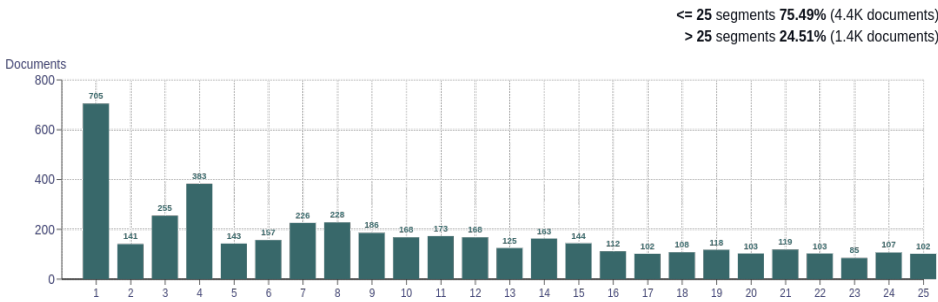
Top 10 domains

Domain	Docs	% of total
jw.org	2.6K	45.19
biblica.com	955	16.30
biblegateway.com	594	10.14
wikipedia.org	298	5.09
bible.is	225	3.84
4jehovah.org	168	2.87
amoafowaa.com	141	2.41
4mormon.org	44	0.75
kasahorow.org	42	0.72
vanipedia.org	31	0.53

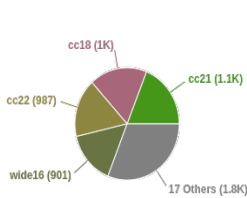
Top 10 TLDs

Domain	Docs	% of total
org	3.5K	58.87
com	2.1K	35.22
is	225	3.84
net	30	0.51
info	9	0.15
de	7	0.12
fm	6	0.10
edu	5	0.09
com.au	5	0.09
cat	4	0.07

Documents size (in segments)

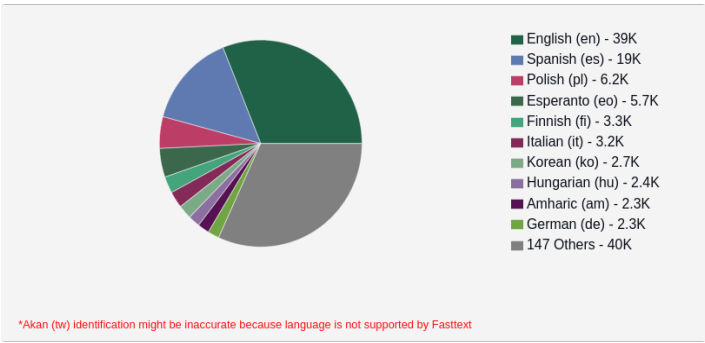


Documents by collection

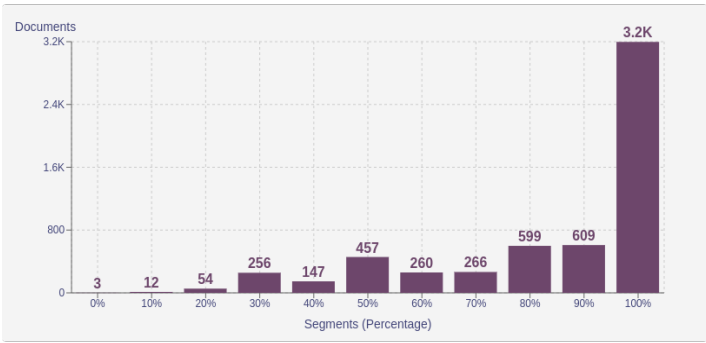


Language Distribution

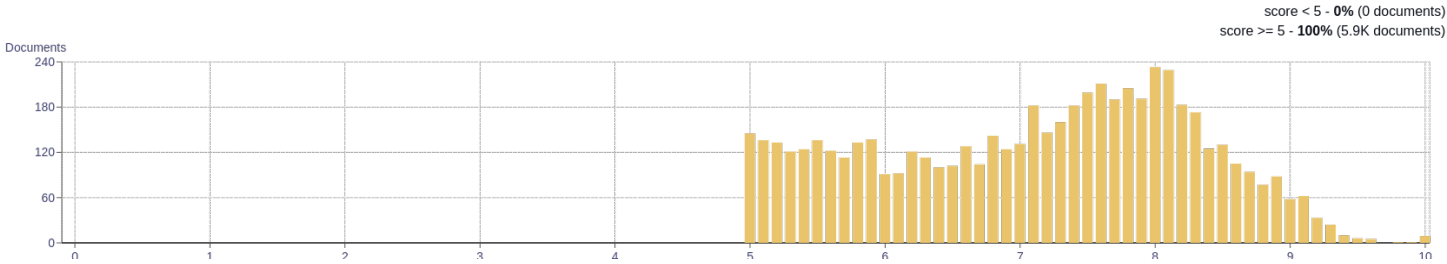
Number of segments



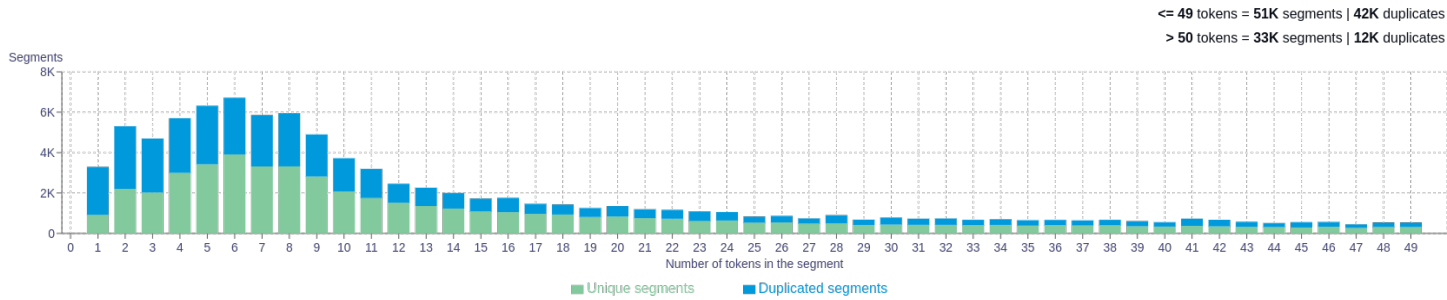
Percentage of segments in Akan (tw) inside documents



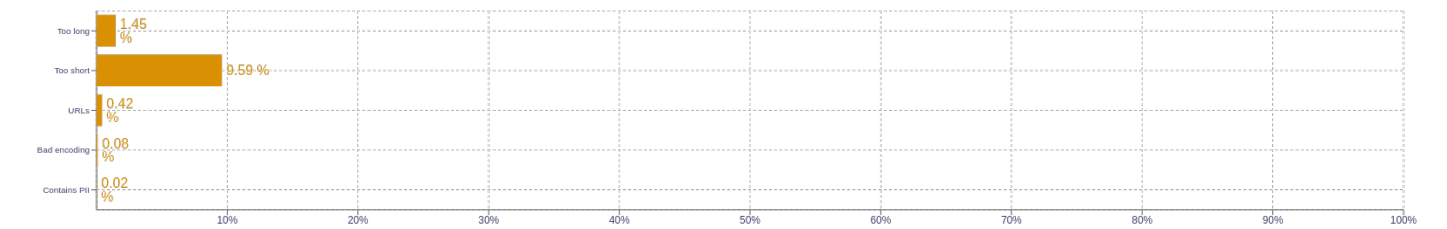
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div><div>mu   64457</div><div>wɔ̯   57424</div><div>wɔ̯n   48169</div><div>wo   31925</div><div>ma   24027</div></div>
2	<div><div>yesu kristo   2100</div><div>yi mu   2056</div><div>yehowa adansefo   1945</div><div>saa ara   1865</div><div>ɔ̯wɔ̯ kɛ   1720</div></div>
3	<div><div>koma no page   1179</div><div>saa asem yi   877</div><div>ase ase ase   832</div><div>kyerɛɛ no se   557</div><div>wɔ̯ mekɛ mɔ̯ɔ̯   525</div></div>
4	<div><div>koma no page desu   1179</div><div>amoros koma no page   1179</div><div>ase ase ase ase   828</div><div>boa boa boa boa   312</div><div>saa asem yi wɔ̯   287</div></div>
5	<div><div>amoros koma no page desu   1179</div><div>ase ase ase ase ase   824</div><div>boa boa boa boa boa   302</div><div>wopɛ se wokenkan saa asem   277</div><div>wokenkan saa asem yi wɔ̯   277</div></div>

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 of 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>