HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ewe_Latn.jsonl.tsv	11/27/2024	Ewe (ee)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,772	143,401	62,285 (43.43 %)	5.1M	22.14 MB	21,178,605

Top 10 domains

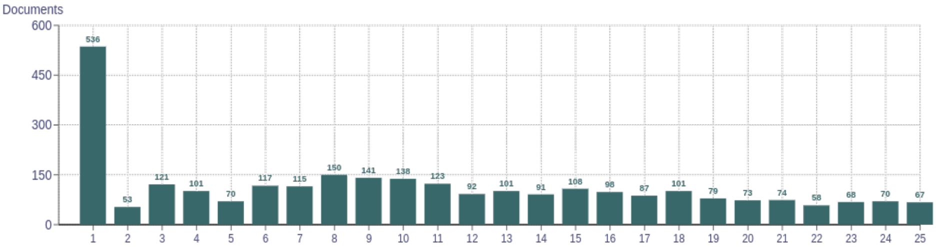
Domain	Docs	% of total
jw.org	2.9K	76.51
wikipedia.org	467	12.38
togochretien.com	86	2.28
mi-eweland.com	45	1.19
bibles.org	30	0.80
voltaonlinegh.com	24	0.64
unicode.org	15	0.40
bible.is	13	0.34
kasahorow.org	13	0.34
ebible.org	13	0.34

Top 10 TLDs

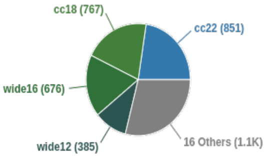
Domain	Docs	% of total
org	3.5K	92.21
com	232	6.15
net	17	0.45
is	13	0.34
info	10	0.27
pl	3	0.08
bible	2	0.05
blog	2	0.05
cn	2	0.05
eu	1	0.03

Documents size (in segments)

<= 25 segments **75.08%** (2.8K documents)
> 25 segments **24.92%** (940 documents)

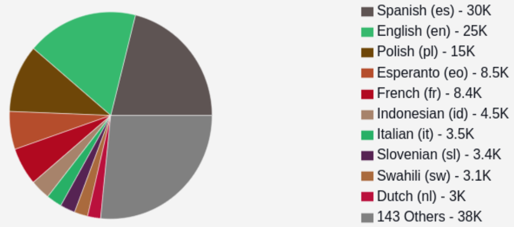


Documents by collection



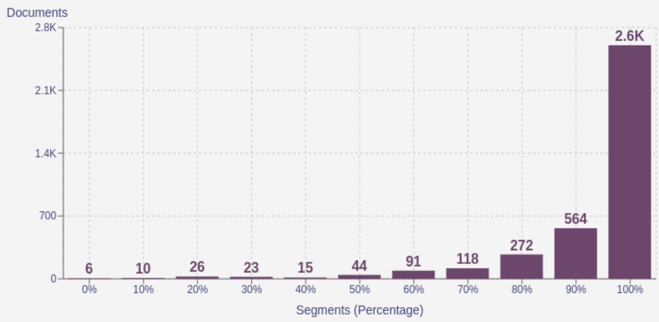
Language Distribution

Number of segments



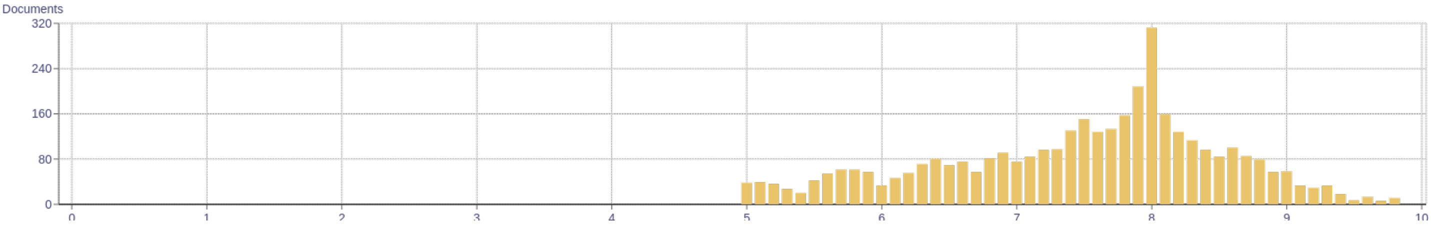
*Ewe (ee) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Ewe (ee) inside documents



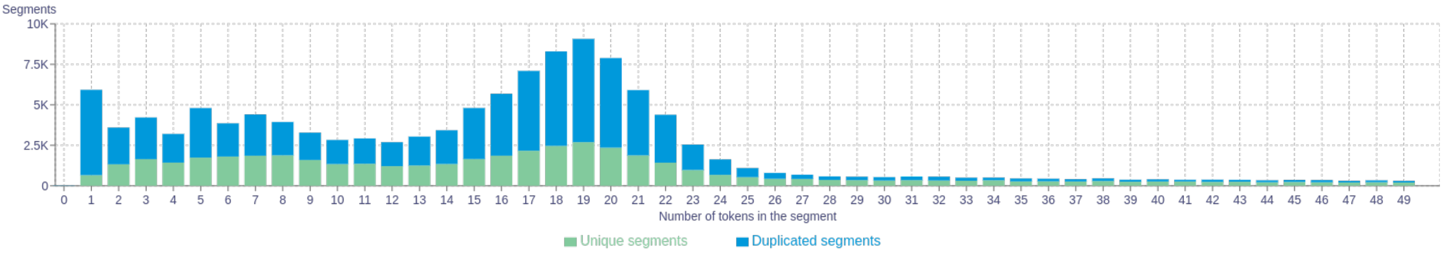
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.8K documents)

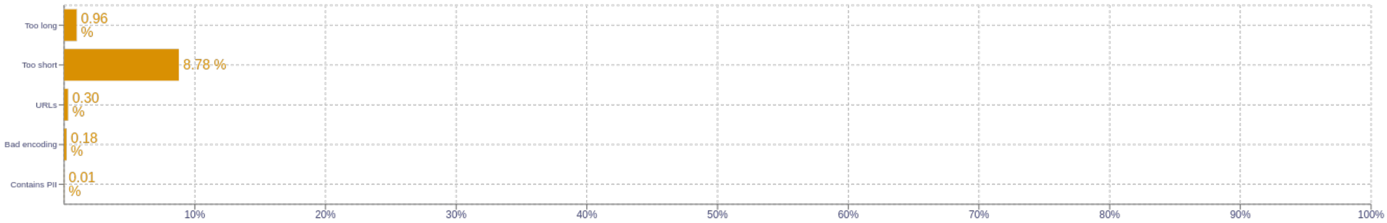


Segment length distribution by token

<= 49 tokens = **46K** segments | **75K** duplicates
> 50 tokens = **22K** segments | **5.7K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	efe | 20710 yehowa | 16739 biblia | 10777 wofe | 9147 aɖeke | 7097
2	yehowa ðasefowo | 1665 gbɔgbɔ kɔkɔe | 1390 sue sue | 1314 efe nusrɔ | 1291 nenema kee | 730
3	sue sue sue | 1306 afetɔ yesu kristo | 438 adam kple xawa | 328 biblia ɖo eŋui | 273 xexea me godoo | 259
4	siasia sia sia siasia | 3325 sue sue sue sue | 1299 trɔ asi le etsofe | 673 nɔnɔmetata si le axa | 406 twj tso kua mis | 200
5	sue sue sue sue sue | 1292 ate ŋu akpe ɖe ŋuwò | 291 tututue nye biblia fe nufiafia | 126 slurry twj tso kua mis | 86 xɔasi siwo le mawu ƒe | 83

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