HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ewe_Latn	9/17/2025	Ewe

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7,137	218,033	196,843 (90.28 %)	9.5M	39,738,732	41.72 MB

Top 10 domains

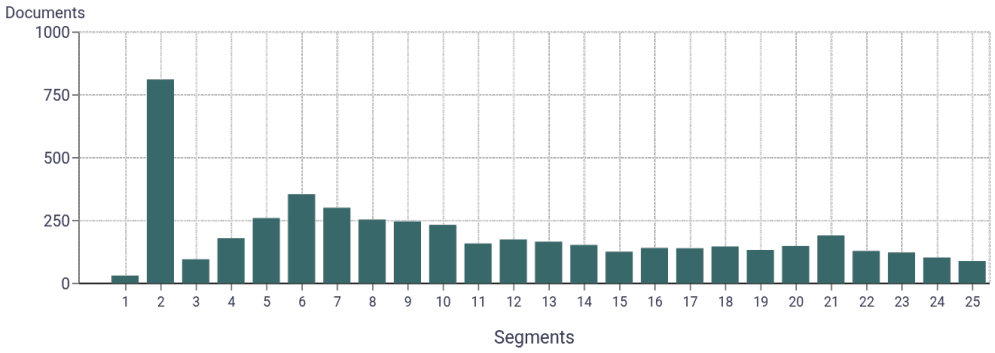
Domain	Docs	% of total
jw.org	5K	70.00%
lyfta.app	910	12.75%
eu5.org	227	3.18%
wikipedia.org	125	1.75%
ebible.org	77	1.08%
degbea.org	60	0.84%
zhitov.ru	57	0.80%
ewelyrics.com	56	0.78%
biblearc.com	42	0.59%
bibles.org	35	0.49%

Top 10 TLDs

Domain	Docs	% of total
org	5.6K	78.51%
app	910	12.75%
com	400	5.60%
ru	60	0.84%
net	33	0.46%
info	33	0.46%
co.uk	26	0.36%
fr	14	0.20%
is	9	0.13%
news	8	0.11%

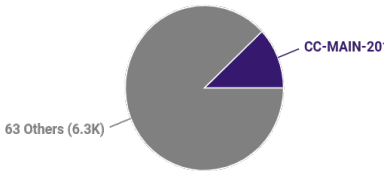
Documents size (in segments)

≤ 25 segments **68.56%** (4.9K documents)
> 25 segments **31.44%** (2.2K documents)



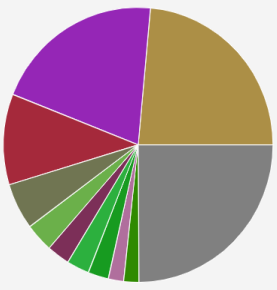
Document collections

CC = 95.84%
IA = 4.16%



Language Distribution

Number of segments in the Ewe corpus

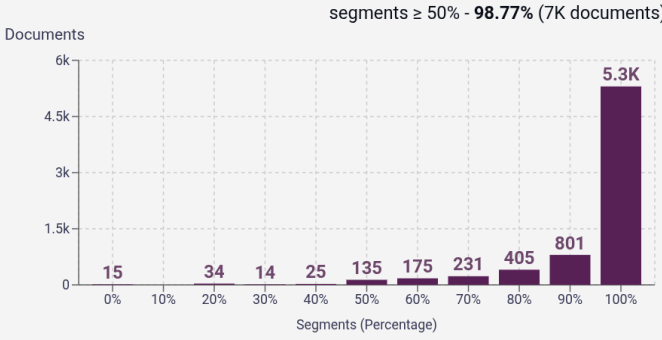


*Ewe identification might be inaccurate because it is not supported by FastSpell.

- Spanish - 51K **(23.5%)**
- English - 45K **(20.5%)**
- Polish - 23K **(10.7%)**
- French - 12K **(5.4%)**
- Esperanto - 7.4K **(3.4%)**
- Italian - 6K **(2.7%)**
- Slovenian - 5.9K **(2.7%)**
- Indonesian - 5.4K **(2.5%)**
- Dutch - 4K **(1.8%)**
- Swahili - 4K **(1.8%)**
- 151 Others - 54K **(24.9%)**

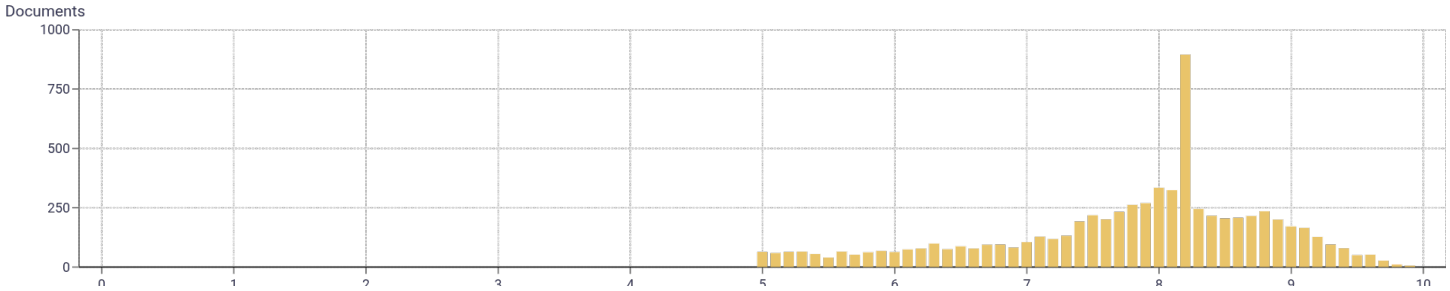
Percentage of segments in Ewe inside documents

segments < 50% - **1.23%** (88 documents)
segments ≥ 50% - **98.77%** (7K documents)



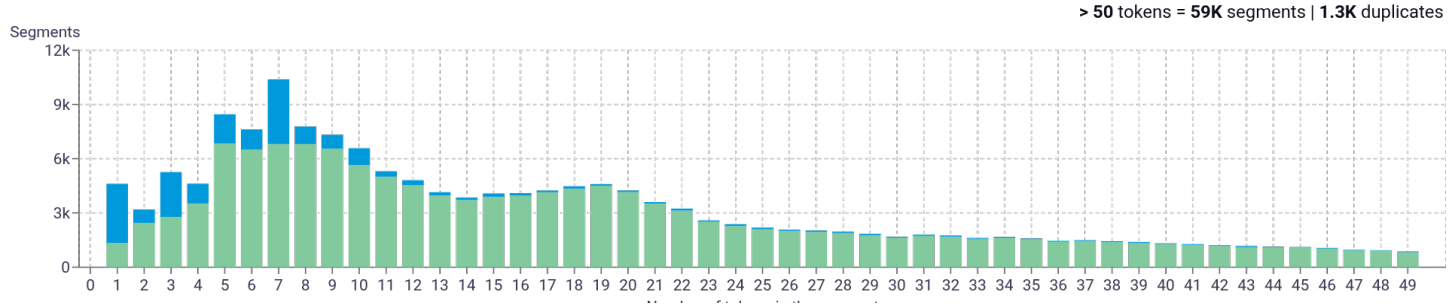
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (7.1K documents)

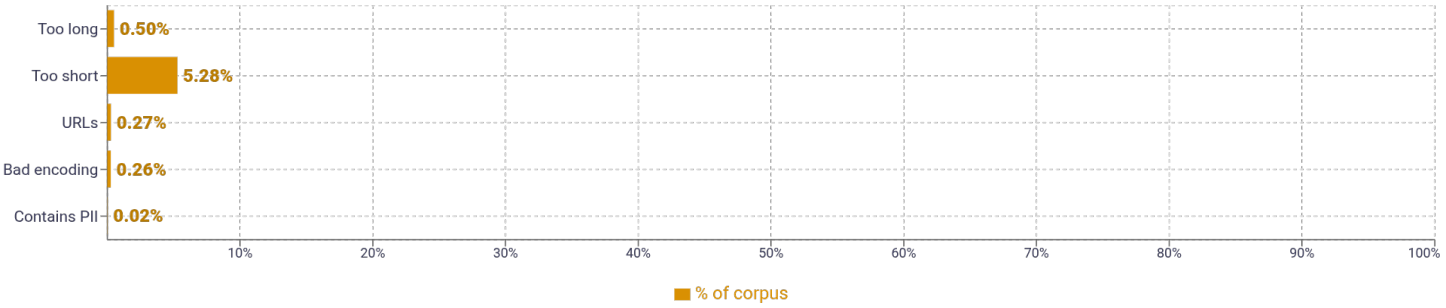


Segment length distribution by token

≤ 49 tokens = **159K** segments | **20K** duplicates
> 50 tokens = **59K** segments | **1.3K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	yehowa | 42,019 · efe | 38,491 · biblia | 20,045 · wofe | 18,500 · ate | 16,477
2	yehowa ðasefowo | 3,297 · in the | 2,511 · akpa gãtɔ | 1,583 · apostolo paulo | 1,401 · nenema kee | 1,253
3	ate ŋu anye | 768 · míate ŋu asrɔ | 733 · ate ŋu akpɔ | 695 · xexe yeye gɔmeɖeɖe | 575 · ate ŋu adzɔ | 563
4	nɔnɔmetata si le axa | 2,197 · nɔnɔmetata siwo le axa | 530 · akpa si le etame | 379 · alesi wòle le kamedede | 352 · fofo si le dzifo | 330
5	ate ŋu akpe ɖe ŋuwò | 502 · alesi wòle le kamedede yeye | 183 · millised on head täiendavad harjutused | 164 · alesi wòle le kamedede ɖesiaɖe | 163 · nɔnɔmetata tsofe si le axa | 152

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				