HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-fon_Latn	9/17/2025	Fon

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,469	24,987	23,712 (94.90 %)	1.9M	6,373,305	7.38 MB

Top 10 domains

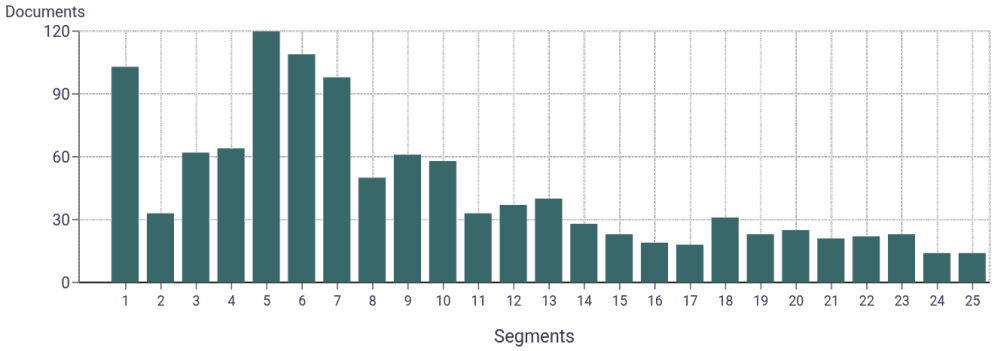
Domain	Docs	% of total
jw.org	1.1K	71.55%
wikipedia.org	182	12.39%
bible.is	90	6.13%
wikimedia.org	66	4.49%
bensino-eg.com	29	1.97%
saxwe.net	15	1.02%
wikiquote.org	4	0.27%
bible.com	4	0.27%
accessagricultu...	4	0.27%
association-ayi...	3	0.20%

Top 10 TLDs

Domain	Docs	% of total
org	1.3K	89.65%
is	90	6.13%
com	40	2.72%
net	18	1.23%
ru	1	0.07%
io	1	0.07%
bj	1	0.07%
bible	1	0.07%

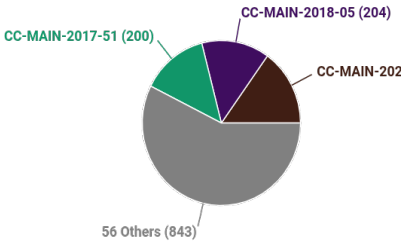
Documents size (in segments)

≤ 25 segments **76.86%** (1.1K documents)
> 25 segments **23.14%** (340 documents)



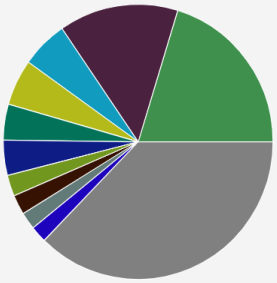
Document collections

CC = 93.67%
IA = 6.33%



Language Distribution

Number of segments in the Fon corpus

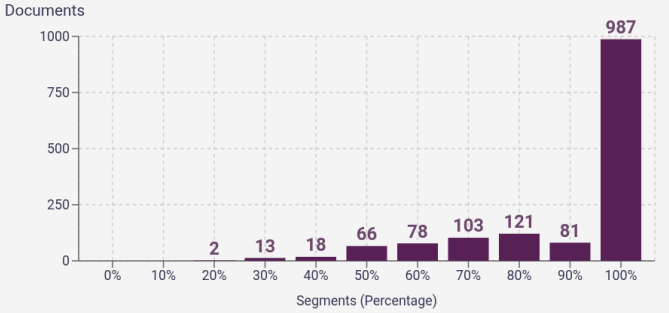


- English - 5.1K **(20.3%)**
- Irish - 3.6K **(14.3%)**
- Spanish - 1.4K **(5.6%)**
- Portuguese - 1.4K **(5.4%)**
- Galician - 1.1K **(4.2%)**
- Czech - 1K **(4.1%)**
- Italian - 620 **(2.5%)**
- Filipino - 582 **(2.3%)**
- Cornish - 505 **(2.0%)**
- Slovenian - 491 **(2.0%)**
- 133 Others - 9.3K **(37.3%)**

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

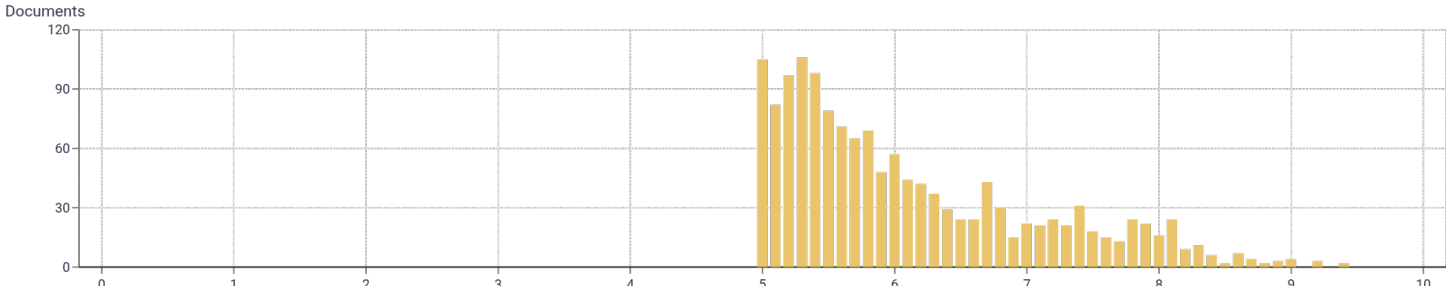
Percentage of segments in Fon inside documents

segments < 50% - **2.25%** (33 documents)
segments ≥ 50% - **97.75%** (1.4K documents)



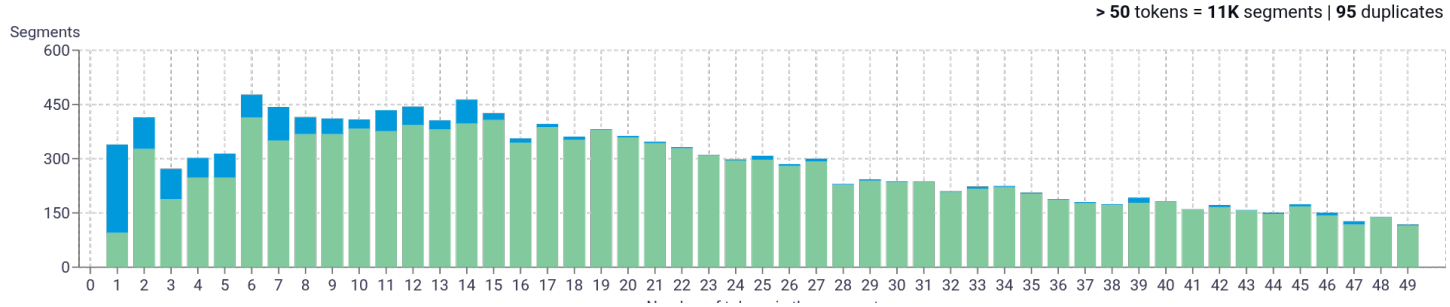
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.5K documents)

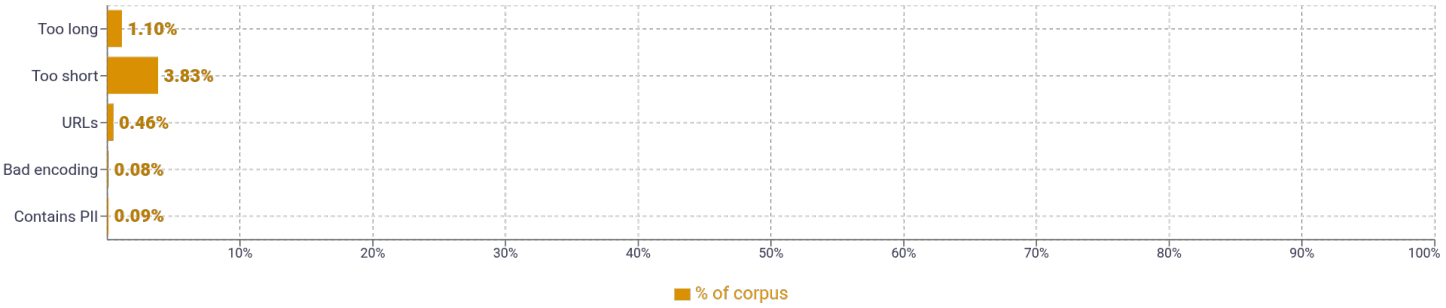


Segment length distribution by token

≤ 49 tokens = **14K** segments | **1.2K** duplicates
> 50 tokens = **11K** segments | **95** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	lɛ | 49,767 · mɛ | 29,335 · na | 28,090 · tɔn | 25,946 · nɛ | 25,278
2	tɔn lɛ | 4,873 · ɖó na | 2,719 · lɛ kpo | 2,442 · mɛ lɛ | 2,053 · mawu tɔn | 1,827
3	mɛ ɖevo lɛ | 947 · lɛ é kpo | 803 · mĭ ɖó na | 688 · wɛ nyí ɖɔ | 606 · yí wăn nú | 605
4	dó wusyɛn lanmɛ nú | 381 · mĭ bɔ mĭ na | 286 · alɔ mĭ bɔ mĭ | 268 · mĭ ka ɖó na | 234 · ɖu ayi mɛ nú | 198
5	alɔ mĭ bɔ mĭ na | 181 · alɔ we bɔ a na | 119 · nɔví sunnu kpo nɔví nyɔnu | 96 · nɛ a pɔ mɛ nu | 94 · lee nŭ nɔ cí nú | 93

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				