

General overview

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

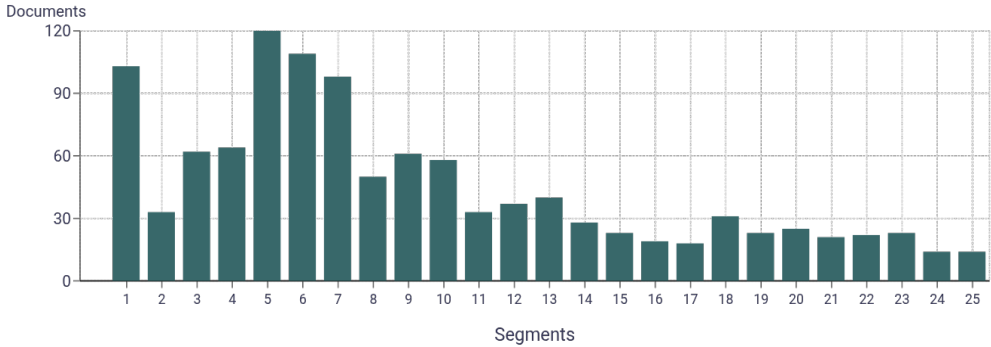
Top 10 domains

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

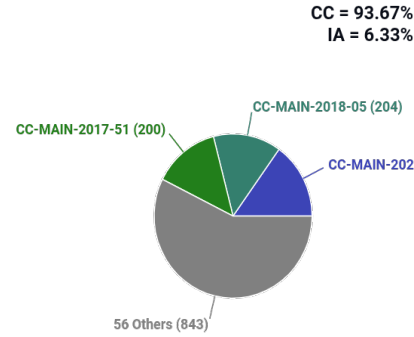
Domain	Docs	% of total
org	1.3K	89.65%
is	90	6.13%
com	40	2.72%
net	18	1.23%
ru	1	0.07%
io	1	0.07%
bj	1	0.07%
bible	1	0.07%

Documents size (in segments) ⓘ



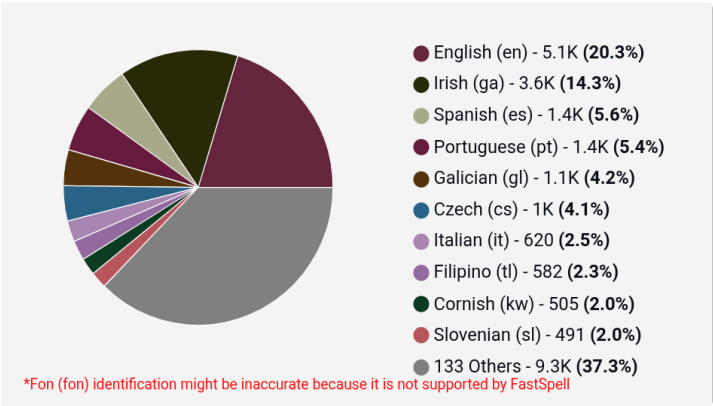
≤ 25 segments **76.86%** (1.1K documents)
> 25 segments **23.14%** (340 documents)

Document collections

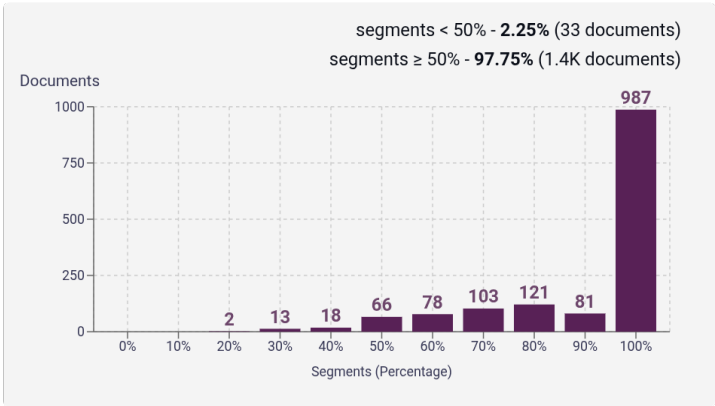


Language Distribution

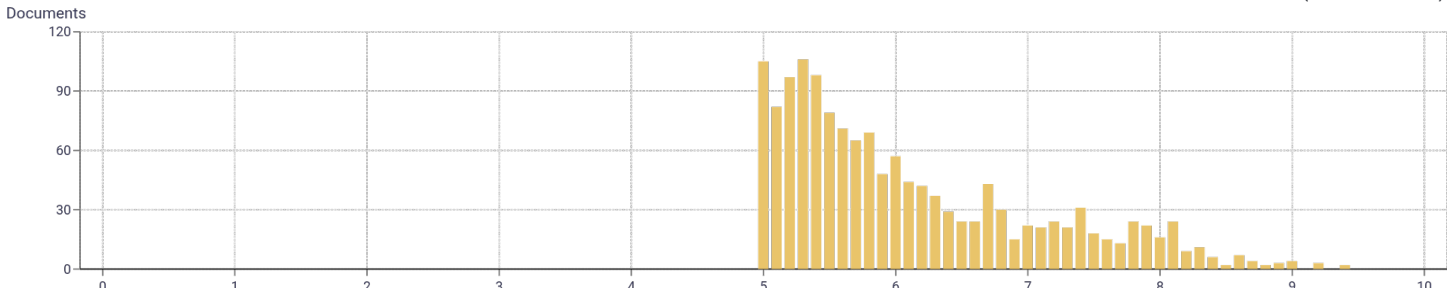
Number of segments in the Fon (fon) corpus



Percentage of segments in Fon (fon) inside documents

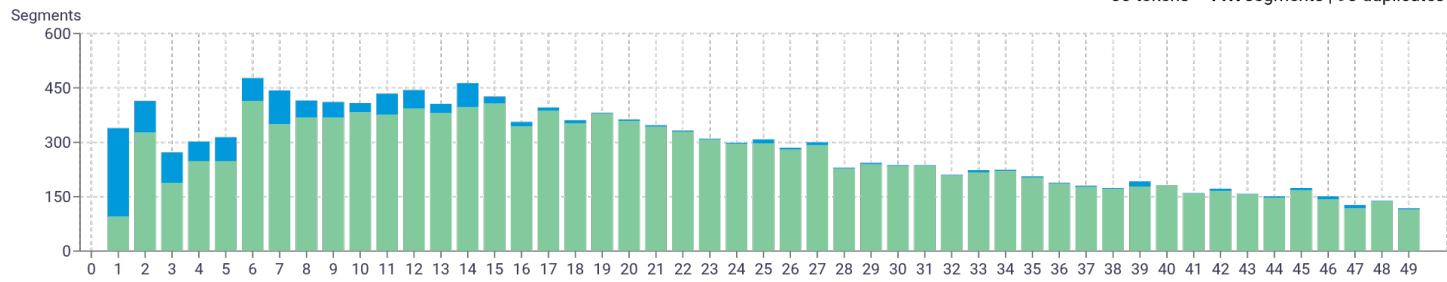


Distribution of documents by document score

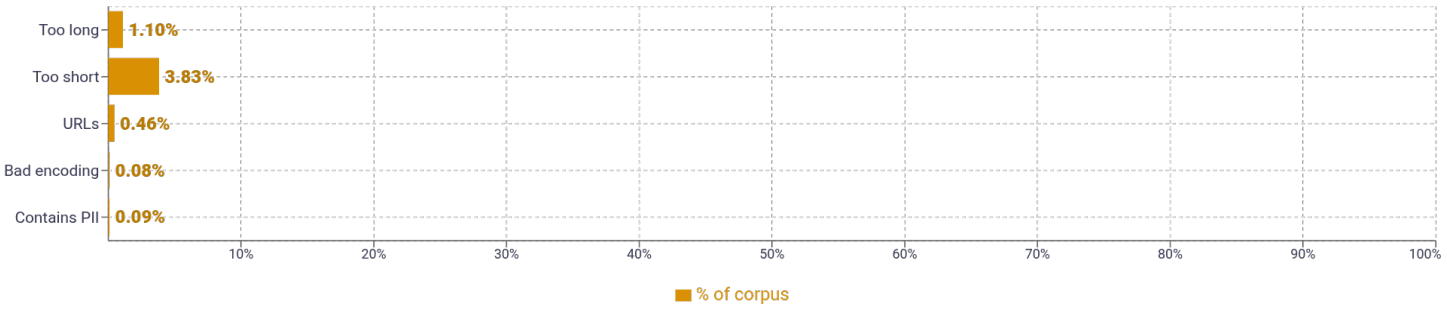


Segment length distribution by token

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



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	le 49,767 me 29,335 na 28,090 tn 25,946 ne 25,278	📄
2	tn le 4,873 qo na 2,719 le kpo 2,442 me le 2,053 mawu tn 1,827	📄
3	me qevo le 947 le e kpo 803 mi qo na 688 we nyi qo 606 yi wān nū 605	📄
4	dó wusyen lanme nū 381 mi bō mi na 286 alō mi bō mi 268 mi ka qo na 234 qu ayi me nū 198	📄
5	alō mi bō mi na 181 alō we bō a na 119 ኦቪ sunnu kpo ኦቪ nyochu 96 ne a pō me nu 94 lee nū nō cí nū 93	📄

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				