HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-kbp_Latn	9/18/2025	Kabiyè (kbp)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,774	68,237	64,474 (94.49 %)	4.9M	19,498,910	23.27 MB

Top 10 domains

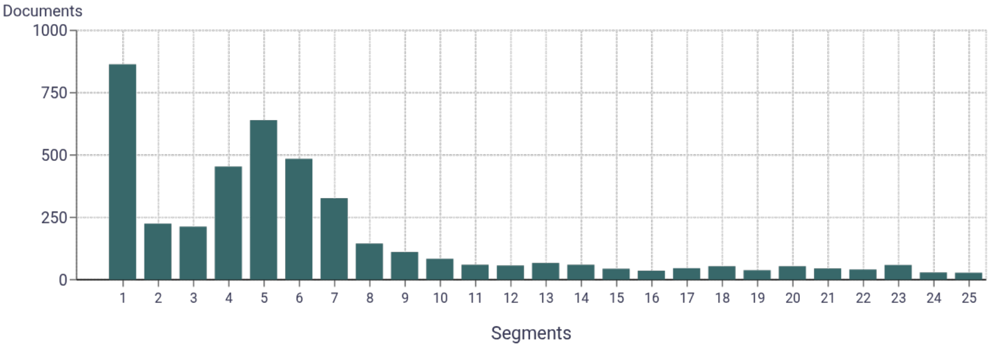
Domain	Docs	% of total
wikipedia.org	1.7K	36.49%
jw.org	1.5K	30.77%
bible.is	986	20.65%
breakeveryyoke.com	213	4.46%
ebible.org	139	2.91%
revue-gugu.org	115	2.41%
wikiplanet.click	33	0.69%
know.cf	16	0.34%
bible.com	9	0.19%
bywiki.com	8	0.17%

Top 10 TLDs

Domain	Docs	% of total
org	3.5K	72.94%
is	986	20.65%
com	248	5.19%
click	33	0.69%
cf	16	0.34%
net	7	0.15%
vn	1	0.02%
de	1	0.02%

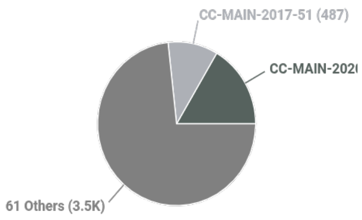
Documents size (in segments) ⓘ

≤ 25 segments **89.36%** (4.3K documents)
> 25 segments **10.64%** (508 documents)



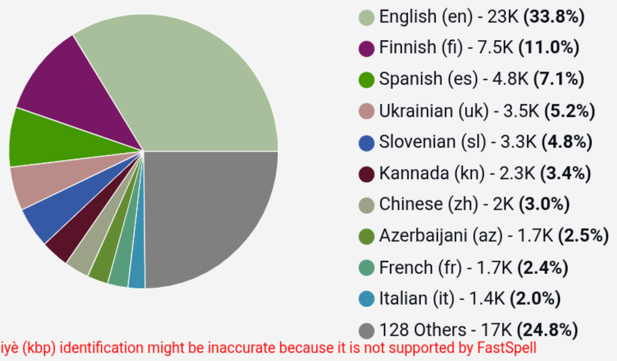
Document collections

CC = 99.18%
IA = 0.82%



Language Distribution

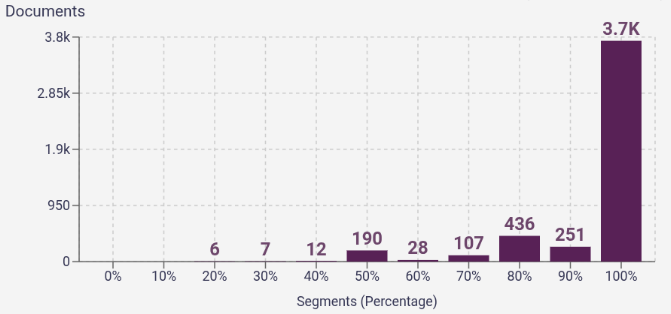
Number of segments in the Kabiyè (kbp) corpus



*Kabiyè (kbp) identification might be inaccurate because it is not supported by FastSpell

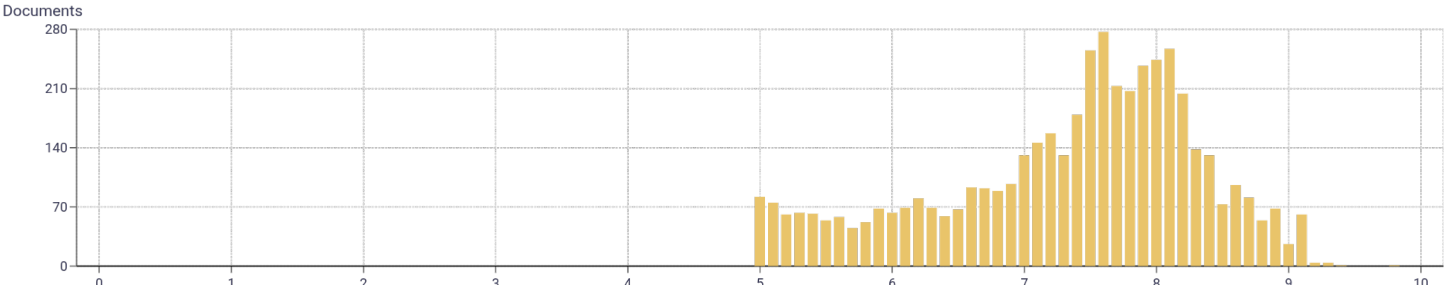
Percentage of segments in Kabiyè (kbp) inside documents

segments < 50% - **0.52%** (25 documents)
segments ≥ 50% - **99.48%** (4.7K documents)



Distribution of documents by document score

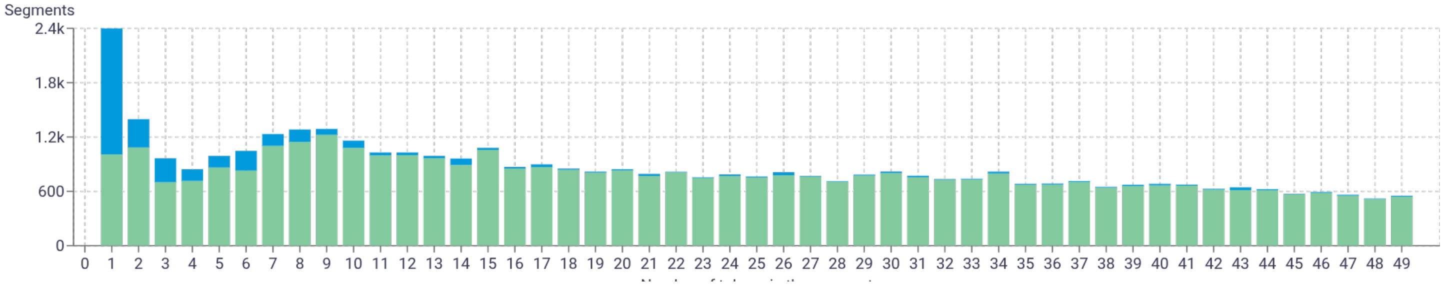
score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4.8K documents)



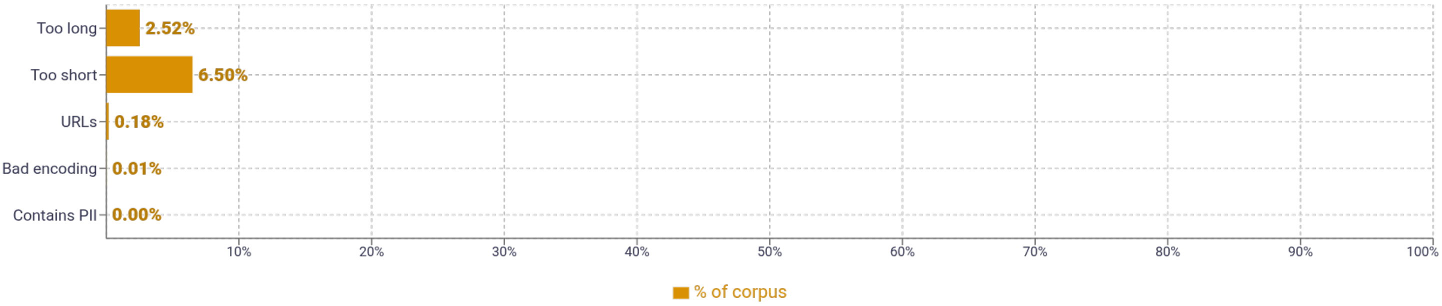
Segment length distribution by token

≤ 49 tokens = **43K** segments | **3.5K** duplicates

> 50 tokens = **25K** segments | **301** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	taa | 75,939 ɩ | 75,067 yɔ | 66,533 se | 59,960 ba | 51,237
2	taa yɔ | 6,990 alɩwaatʊ ndʊ | 4,810 hʊ aa | 4,223 tɔm ndʊ | 4,167 dɩ ba | 3,663
3	mbʊ pʊyɔɔ yɔ | 2,389 nala hʊ aa | 1,257 pə taɣa pʊlʊ | 1,078 hʊ buloŋ aa | 798 pɩtʊʊ fɛyɩ se | 782
4	pə taɣa pʊlʊ tɔɔ | 493 tɩya ba a baa | 478 tɩya ʊ a baa | 399 rɛ yesu basɩ tɩya | 286 wɩsɩ wiyesi welii hʊ | 279
5	basɩ tɩya ba a baa | 448 basɩ tɩya ʊ a baa | 395 kalʊ na ba ti tɩn | 235 hʊ aa bɩ yaa gyuuma | 213 nala hʊ aa bɩ yaa | 209

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				