HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-hat_Latn	9/18/2025	Haitian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
377,114	7,867,052	5,810,563 (73.86 %)	258M	1,178,562,030	1.13 GB

Top 10 domains

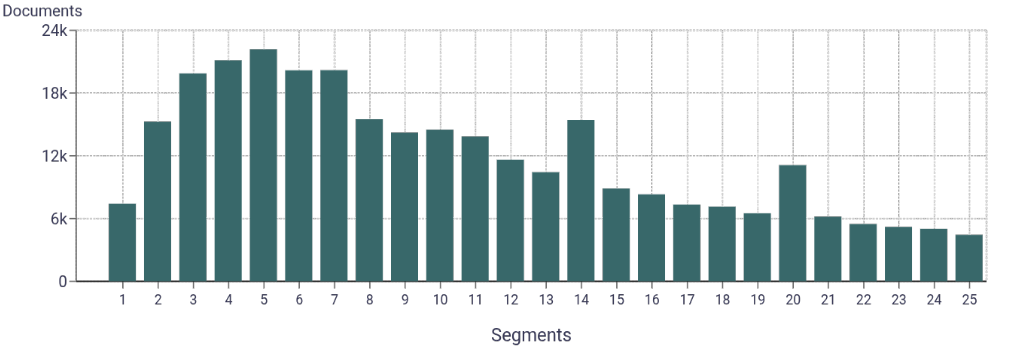
Domain	Docs	% of total
voanouvel.com	16K	4.36%
martech.zone	6.8K	1.80%
itsmygame.org	6.7K	1.77%
socialdesignmag...	6.3K	1.66%
wikipedia.org	4.9K	1.31%
temoignages.re	4.8K	1.27%
eturbonews.com	4.7K	1.25%
wondershare.com	4.7K	1.25%
makeoverarcade.com	4.5K	1.19%
jw.org	3.6K	0.94%

Top 10 TLDs

Domain	Docs	% of total
com	248K	65.64%
org	49K	13.06%
net	15K	3.95%
gov	7.9K	2.09%
zone	6.8K	1.80%
news	5.2K	1.37%
re	4.8K	1.28%
info	3.4K	0.91%
co	2.9K	0.76%
ru	2.5K	0.66%

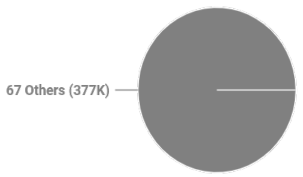
Documents size (in segments) ⓘ

≤ 25 segments **78.89%** (298K documents)
> 25 segments **21.11%** (80K documents)



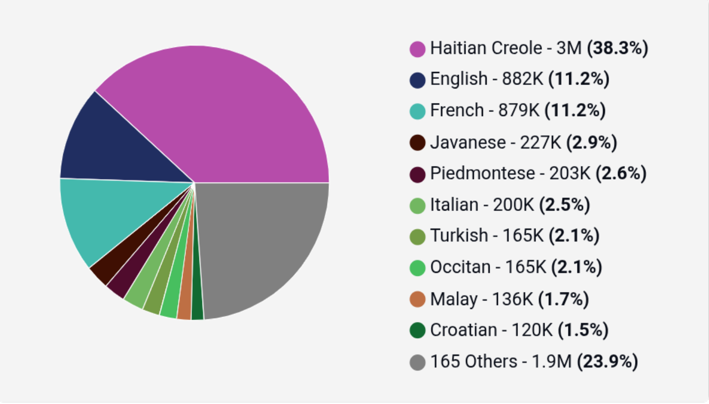
Document collections

CC = 86.96%
IA = 13.04%



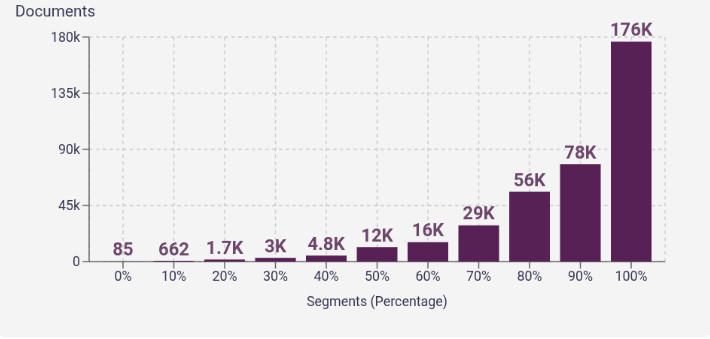
Language Distribution

Number of segments in the Haitian corpus



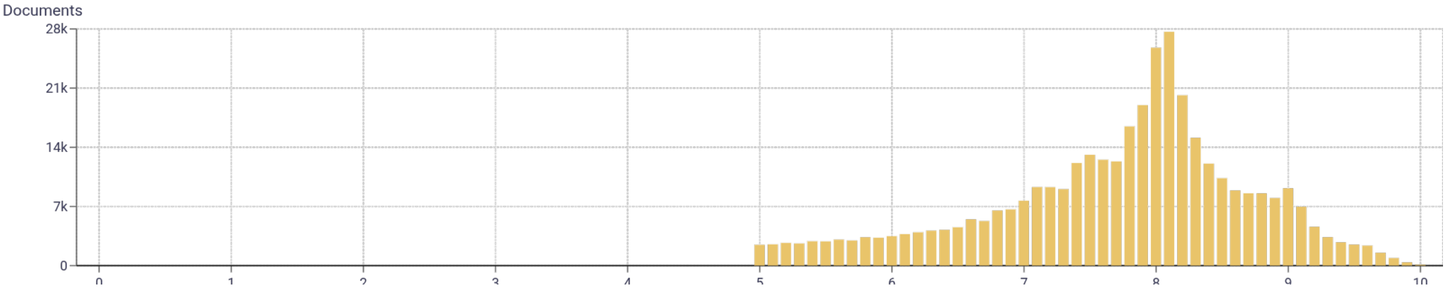
Percentage of segments in Haitian inside documents

segments < 50% - **2.73%** (10K documents)
segments ≥ 50% - **97.27%** (367K documents)



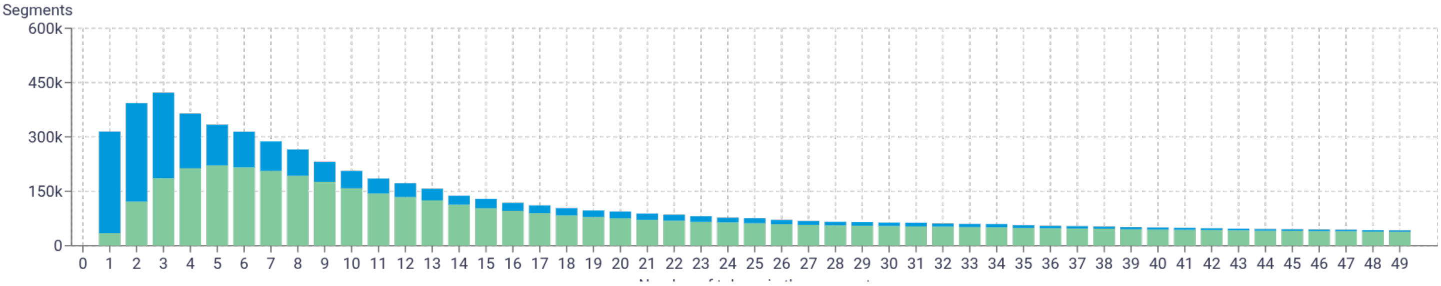
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (377K documents)

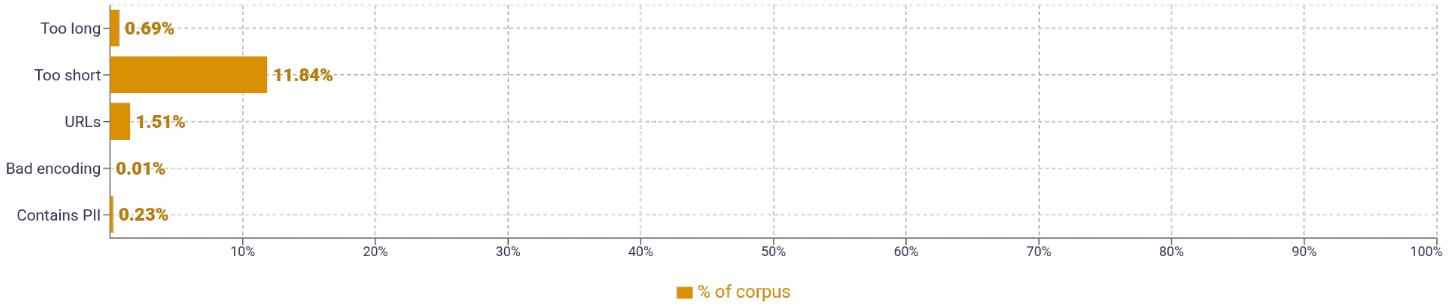


Segment length distribution by token

≤ 49 tokens = **6.2M** segments | **1.9M** duplicates
> 50 tokens = **1.7M** segments | **145K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	moun | 1,025,301 ti | 667,995 bon | 624,043 bay | 458,890 bann | 455,925
2	mr speaker | 116,539 sit entènèt | 86,238 ti fi | 45,880 ti kras | 34,969 divès kalite | 34,000
3	bon jan kalite | 83,760 mersi mr speaker | 20,843 fin vye granmoun | 19,956 mr deputy speaker | 17,434 ti a kontan | 12,466
4	mon ti a kontan | 10,707 bon jan kalite pwodwi | 8,181 pyès ki nan konpitè | 7,131 gwo twou san fon | 6,928 premye a fè kòmantè | 6,524
5	lyen ki nan yon zanmi | 6,102 pataje jwèt la ak mond | 6,101 kòd la html nan sit | 6,101 kopi kòd la ak keratin | 6,101 keratin nan kòd la html | 6,101

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				