HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-plt_Latn	9/18/2025	Plateau Malagasy

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
365,680	6,751,211	4,801,957 (71.13 %)	246M	1,233,297,031	1.16 GB

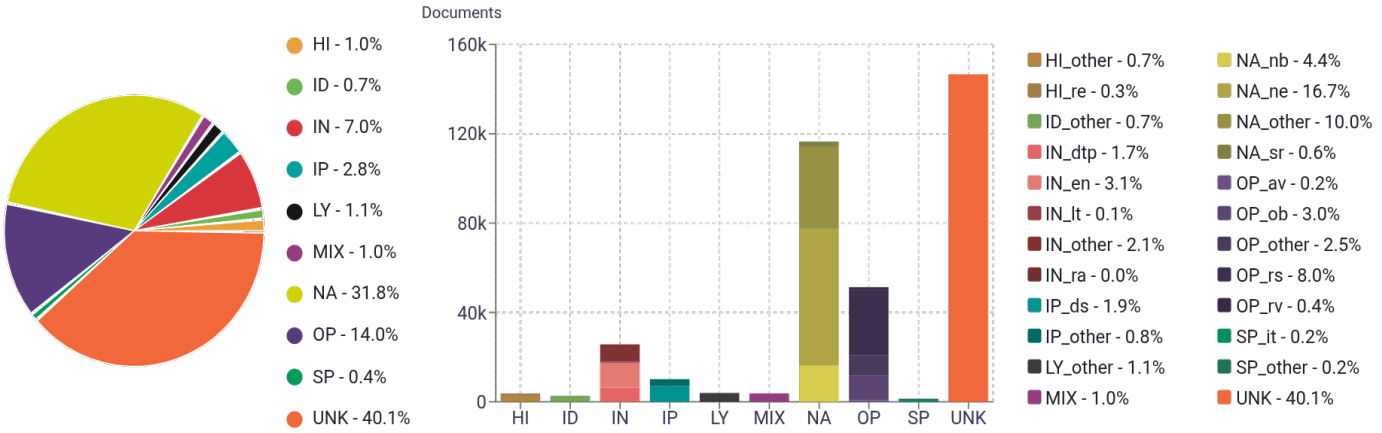
Top 10 domains

Domain	Docs	% of total
globalvoices.org	59K	16.14%
newsmada.com	11K	2.87%
wikipedia.org	9K	2.45%
eturbonews.com	8.6K	2.36%
globalvoicesonl...	8.3K	2.27%
tiatanindrazana.mg	7.7K	2.11%
tiatanindrazana...	6.6K	1.80%
midi-madagasika...	5.9K	1.61%
blaogy.com	5.8K	1.58%
serasera.org	5.4K	1.48%

Top 10 TLDs

Domain	Docs	% of total
com	175K	47.76%
org	113K	30.81%
mg	30K	8.22%
net	11K	2.92%
zone	4.7K	1.29%
info	4.6K	1.25%
gov.mg	3.1K	0.84%
news	2.7K	0.73%
fr	2.5K	0.70%
es	1.5K	0.42%

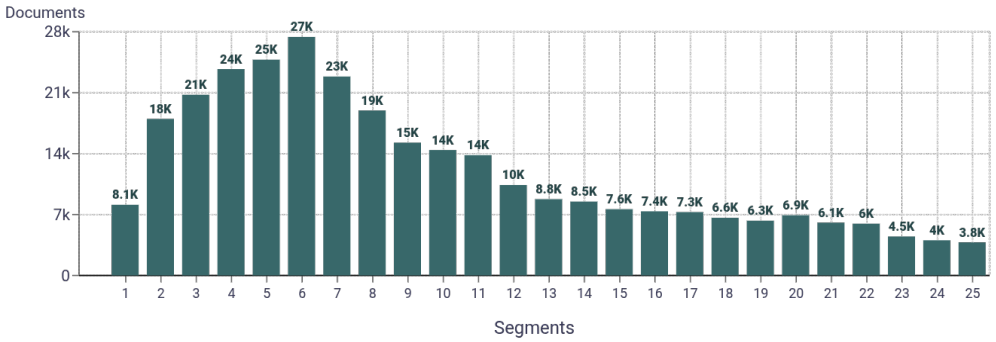
Register labels



MT:33.9% | 124K Documents

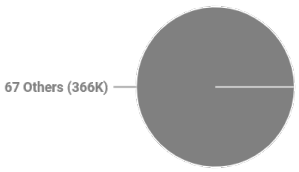
Documents size (in segments)

≤ 25 segments **82.61%** (302K documents)
> 25 segments **17.39%** (64K documents)



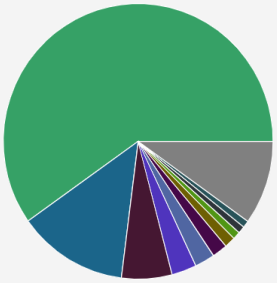
Document collections

CC = 92.51%
IA = 7.49%



Language Distribution

Number of segments in the Plateau Malagasy corpus

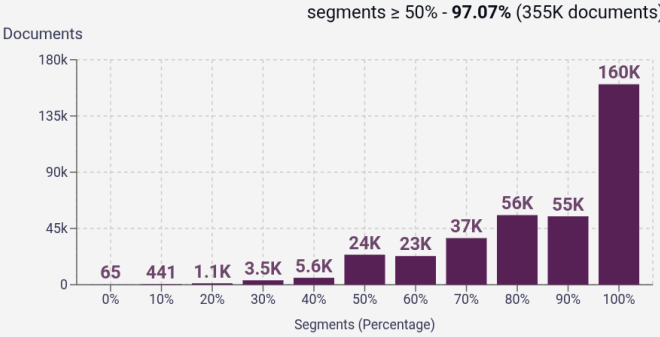


*Plateau Malagasy identification might be inaccurate because it is not supported by FastSpell.

- Malagasy - 4M **(59.8%)**
- English - 891K **(13.2%)**
- French - 409K **(6.1%)**
- Spanish - 201K **(3.0%)**
- Italian - 162K **(2.4%)**
- Basque - 123K **(1.8%)**
- Dutch - 88K **(1.3%)**
- Polish - 61K **(0.9%)**
- Swahili - 56K **(0.8%)**
- German - 56K **(0.8%)**
- 163 Others - 666K **(9.9%)**

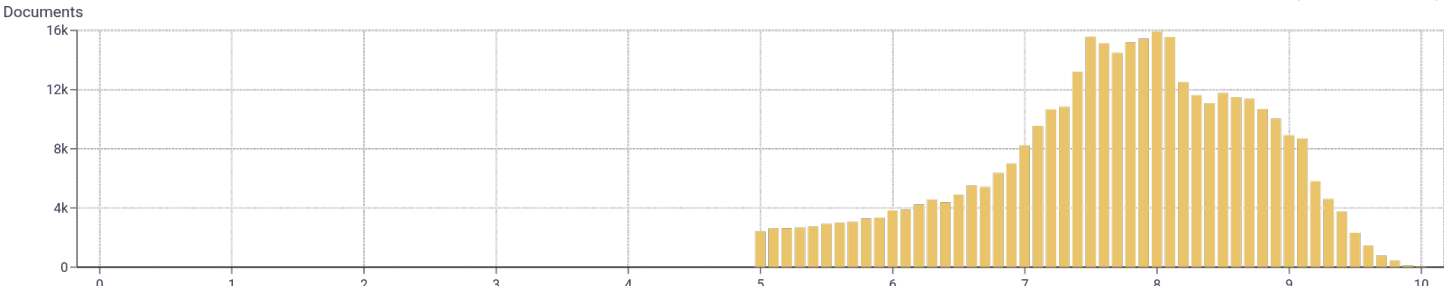
Percentage of segments in Plateau Malagasy inside documents

segments < 50% - **2.93%** (11K documents)
segments ≥ 50% - **97.07%** (355K documents)



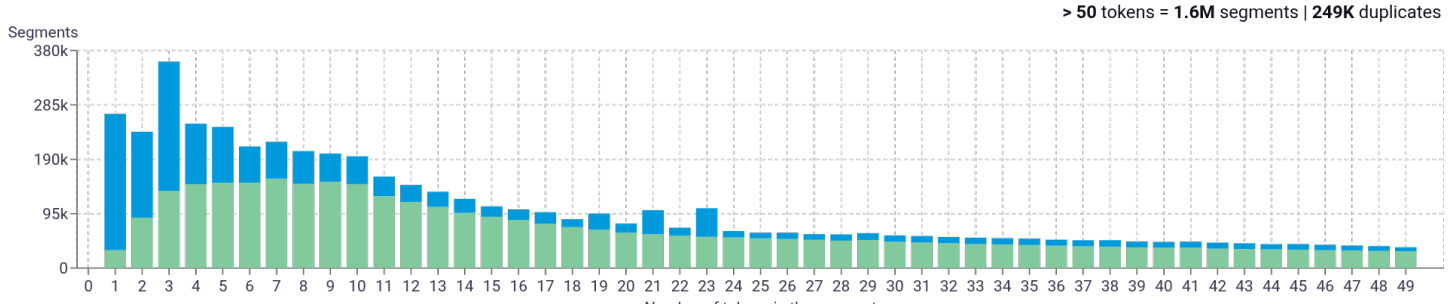
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (366K documents)

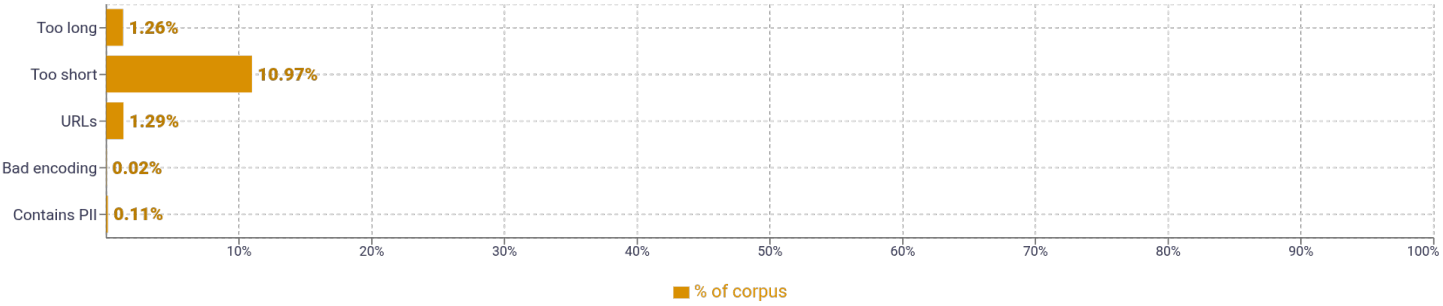


Segment length distribution by token

≤ 49 tokens = **5.1M** segments | **1.7M** duplicates
> 50 tokens = **1.6M** segments | **249K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	amin | 6,003,588 · dia | 4,241,529 · tsy | 2,764,840 · sy | 2,744,664 · ireo | 2,216,743
2	izy ireo | 372,744 · avy amin | 289,528 · dia tsy | 289,371 · eo amin | 269,378 · miaraka amin | 254,338
3	izao tontolo izao | 78,164 · izao fotoana izao | 62,497 · na izany aza | 54,983 · dia tsy maintsy | 47,306 · be dia be | 38,135
4	na inona na inona | 34,353 · koa rehefa mieritreritra hoe | 28,166 · mieritreritra hoe banky dia | 28,165 · izaho koa rehefa mieritreritra | 28,165 · hoe banky dia ireo | 28,165
5	mieritreritra hoe banky dia ireo | 28,165 · koa rehefa mieritreritra hoe banky | 28,165 · izaho koa rehefa mieritreritra hoe | 28,165 · hoe banky dia ireo bankin | 28,165 · tsy maintsy mahay manoratra amin | 16,548

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				