HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
plt_Latn.jsonl.tsv	12/3/2024	Malagasy (plt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
207,837	4,736,104	2,307,265 (48.72 %)	162M	805,769,531	781.36 MB

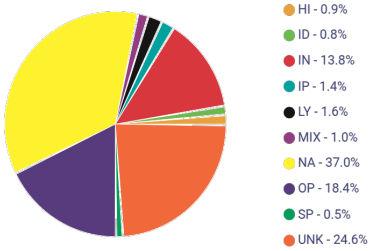
Top 10 domains

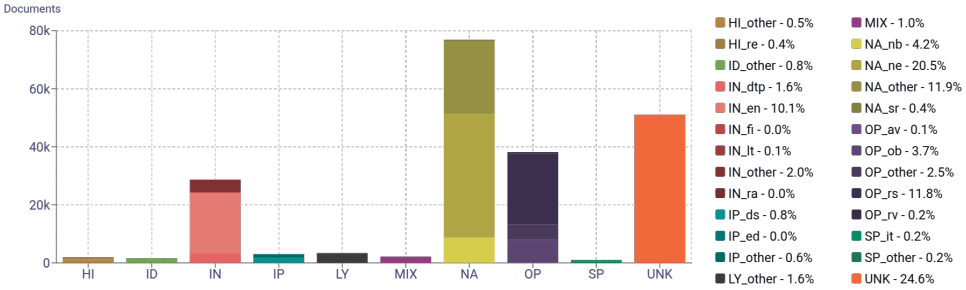
Domain	Docs	% of total
globalvoices.org	47K	22.64%
globalvoicesonl...	18K	8.72%
wikipedia.org	12K	5.60%
wiktionary.org	10K	5.03%
blaogy.com	8.7K	4.20%
katolika.org	6.9K	3.30%
jw.org	5.8K	2.78%
mydago.com	4.2K	2.04%
serasera.org	3.9K	1.88%
tiatanindrazana...	2.5K	1.21%

Top 10 TLDs

Domain	Docs	% of total
org	117K	56.34%
com	64K	30.63%
mg	6.9K	3.31%
net	6K	2.90%
info	2.6K	1.27%
fr	1.8K	0.85%
news	1.4K	0.66%
zone	1K	0.49%
gov.mg	1K	0.48%
is	718	0.35%

Register labels

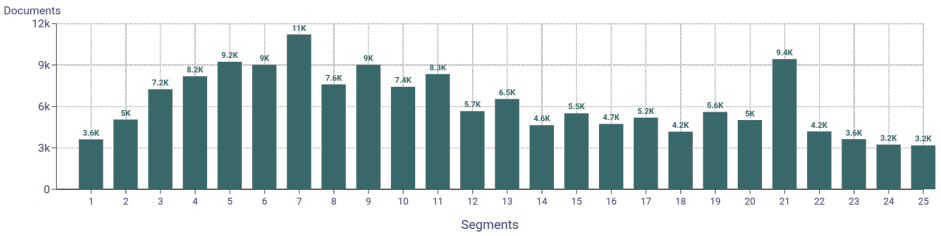




MT:18.7% | 39K Documents

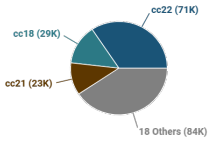
Documents size (in segments)

<= 25 segments **75.34%** (157K documents)
> 25 segments **24.66%** (51K documents)



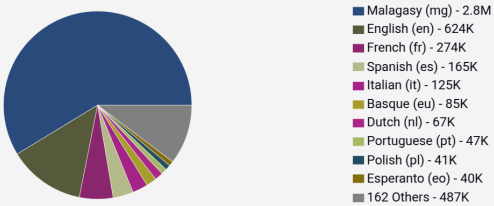
Documents by collection

CC = 71.69%
IA = 28.31%



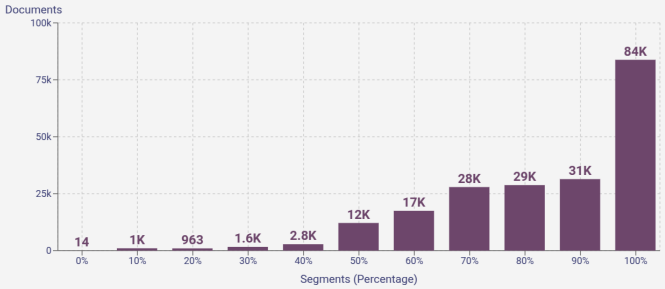
Language Distribution

Number of segments in the Malagasy (plt) corpus



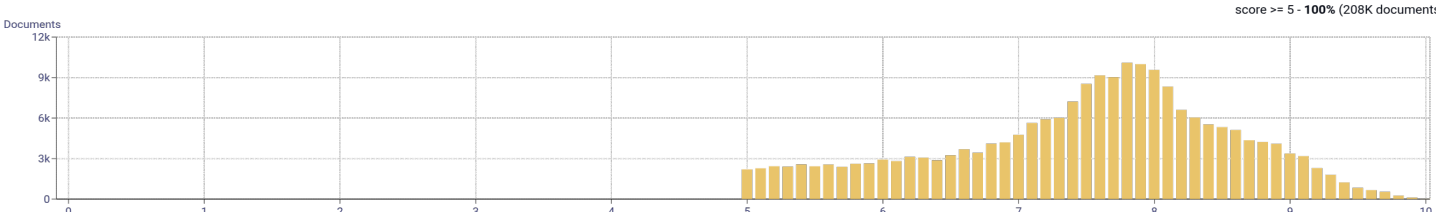
*Malagasy (plt) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Malagasy (plt) inside documents



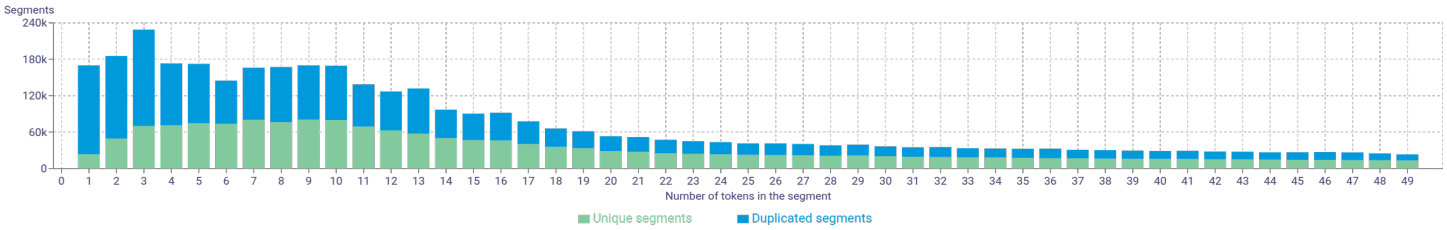
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (208K documents)

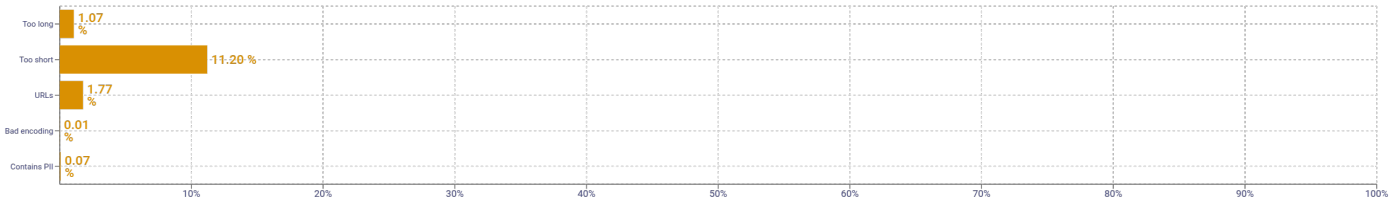


Segment length distribution by token

≤ 49 tokens = **1.7M** segments | **2M** duplicates
> 50 tokens = **1.1M** segments | **446K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	amin | 3513070 · dia | 2224112 · tsy | 1854019 · sy | 1635454 · ireo | 1598717
2	izy ireo | 226318 · avy amin | 201703 · eo amin | 165098 · ihany koa | 153624 · dia tsy | 153264
3	izao tontolo izao | 50675 · izao fotoana izao | 37732 · tsy ampy amin | 36044 · zavatra tsy ampy | 34284 · hanova ny fango | 32097
4	fanononana tsy ampy amin | 26228 · na inona na inona | 20307 · tokony homarinana avy amin | 10479 · dikanteny tokony homarinana avy | 10449 · manerana izao tontolo izao | 8354
5	dikanteny tokony homarinana avy amin | 10449 · araka ny tokony ho izy | 3681 · koa ve ity lahatsoratra nivoaka | 3466 · tadidinao koa ve ity lahatsoratra | 3465 · ve ity lahatsoratra nivoaka tato | 3464

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av