

General overview

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

Volumes

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

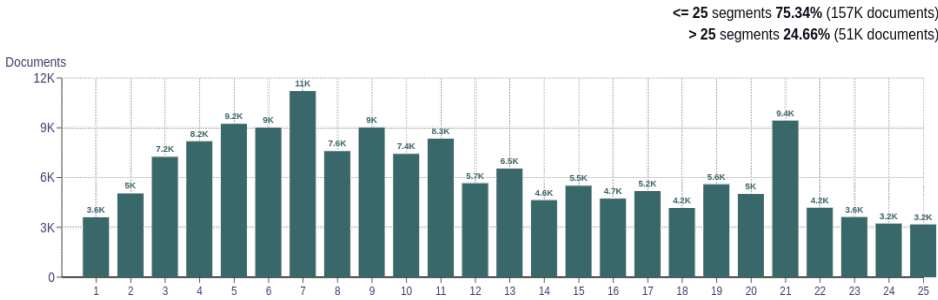
Top 10 domains

Domain	Docs	% of total
globalvoices.org	47K	22.64
globalvoicesonline.org	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.com	2.5K	1.21

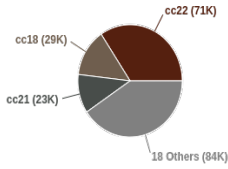
Top 10 TLDs

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

Documents size (in segments)

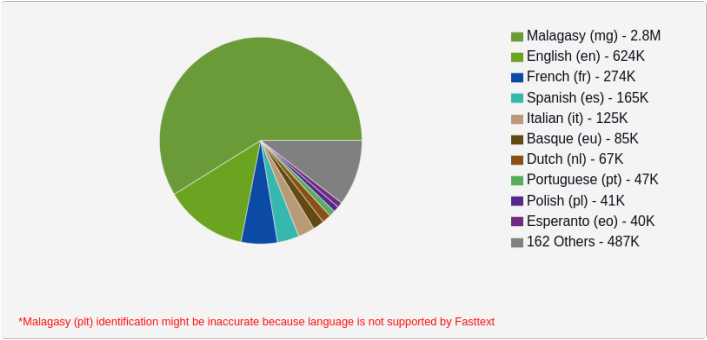


Documents by collection

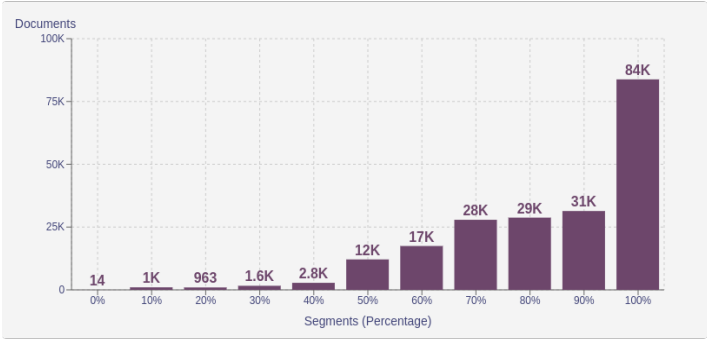


Language Distribution

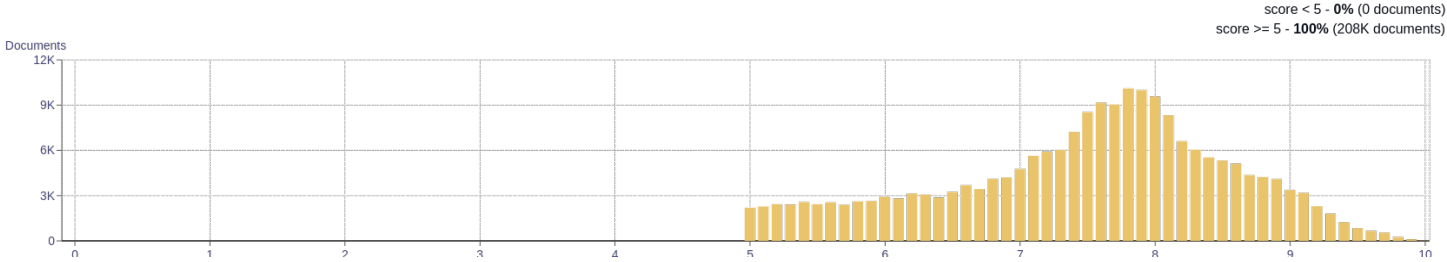
Number of segments



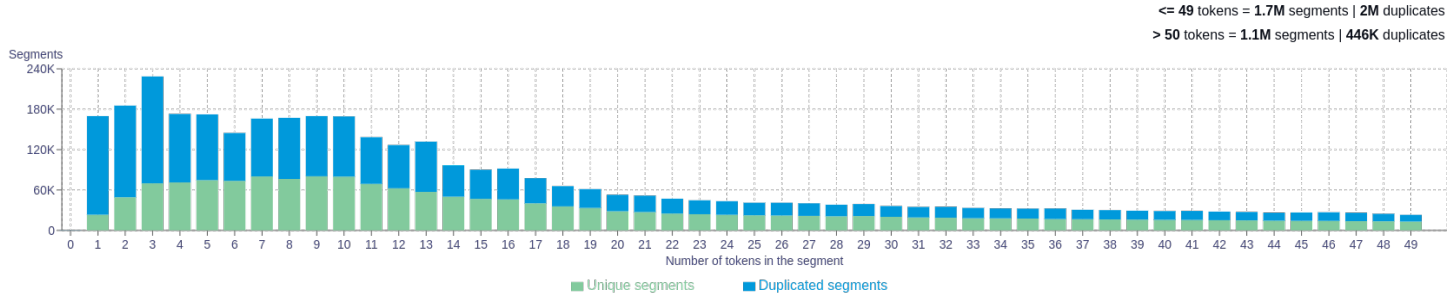
Percentage of segments in Malagasy (plt) inside documents



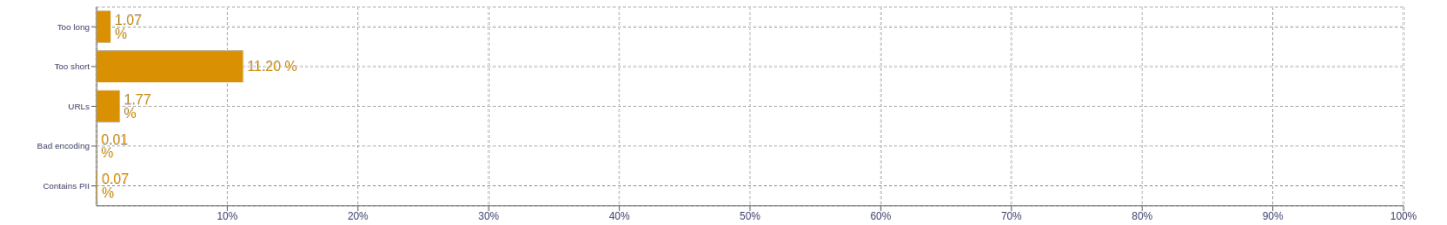
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>amin   3513070</div> <div>dia   2224112</div> <div>tsy   1854019</div> <div>sy   1635454</div> <div>ireo   1598717</div>
2	<div>izy ireo   226318</div> <div>avy amin   201703</div> <div>eo amin   165098</div> <div>ihany koa   153624</div> <div>dia tsy   153264</div>
3	<div>izao tontolo izao   56675</div> <div>izao fotoana izao   37732</div> <div>tsy ampy amin   36044</div> <div>zavatra tsy ampy   34284</div> <div>hanova ny fango   32097</div>
4	<div>fanononana tsy ampy amin   26228</div> <div>na inona na inona   20307</div> <div>tokony homarinana avy amin   10479</div> <div>dikanten'ny tokony homarinana avy   10449</div> <div>manerana izao tontolo izao   8354</div>
5	<div>dikanten'ny tokony homarinana avy amin   10449</div> <div>araka ny tokony ho izy   3681</div> <div>koa ve ity lahatsoratra nivoaka   3466</div> <div>tadidinao koa ve ity lahatsoratra   3465</div> <div>ve ity lahatsoratra nivoaka tato   3464</div>

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 of 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>