HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tir_Ethi.jsonl.tsv	9/19/2024	Tigrinya (ti)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
64,689	1,128,087	599,237 (53.12 %)	43M	180,568,832	434.15 MB

Top 10 domains

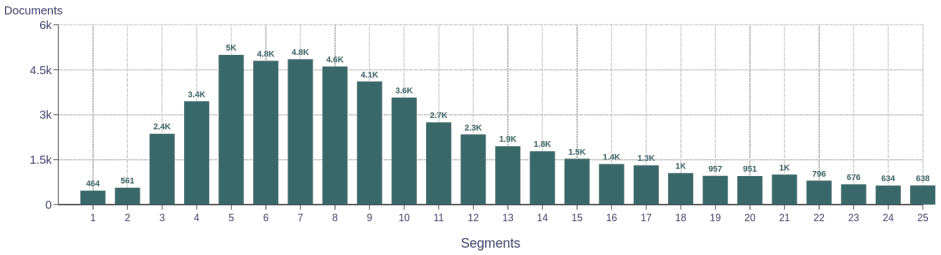
Domain	Docs	% of total
voanews.com	12K	18.87
assenna.com	12K	17.83
erena.org	7K	10.79
vaticannews.va	2.5K	3.86
jw.org	1.9K	3.00
farajat.net	1.5K	2.32
asmarino.com	1.4K	2.24
harnnet.org	1.2K	1.89
informationsverige.se	955	1.48
bbc.com	944	1.46

Top 10 TLDs

Domain	Docs	% of total
com	36K	55.99
org	16K	24.33
va	3K	4.68
se	2.8K	4.28
net	2.2K	3.33
is	854	1.32
de	759	1.17
ch	614	0.95
no	608	0.94
info	548	0.85

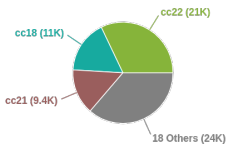
Documents size (in segments)

<= 25 segments **82.61%** (53K documents)
> 25 segments **17.39%** (11K documents)



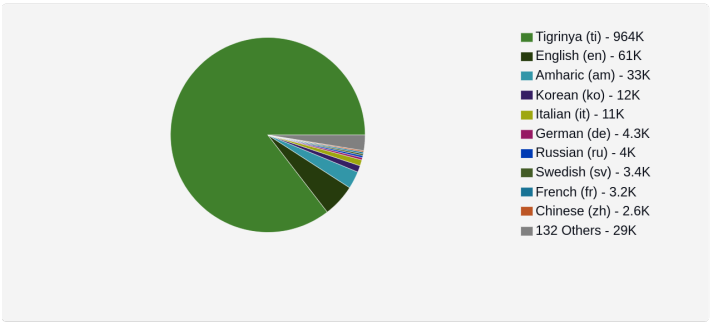
Documents by collection

CC = 72.01%
IA = 27.99%

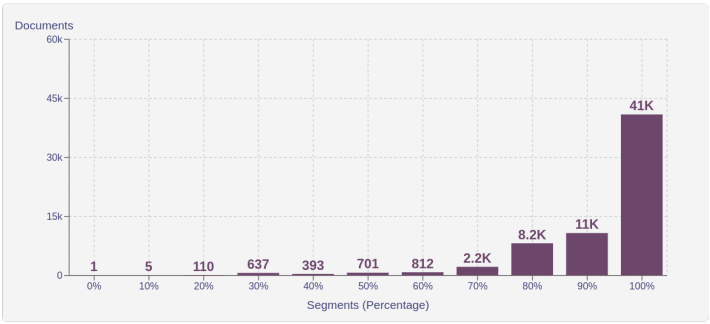


Language Distribution

Number of segments in the Tigrinya (ti) corpus

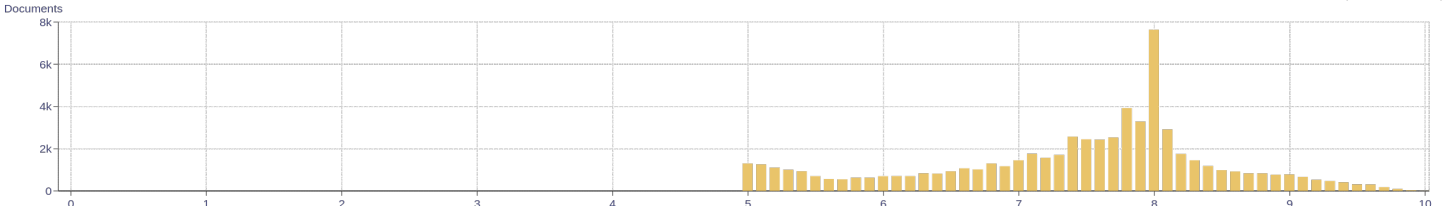


Percentage of segments in Tigrinya (ti) inside documents



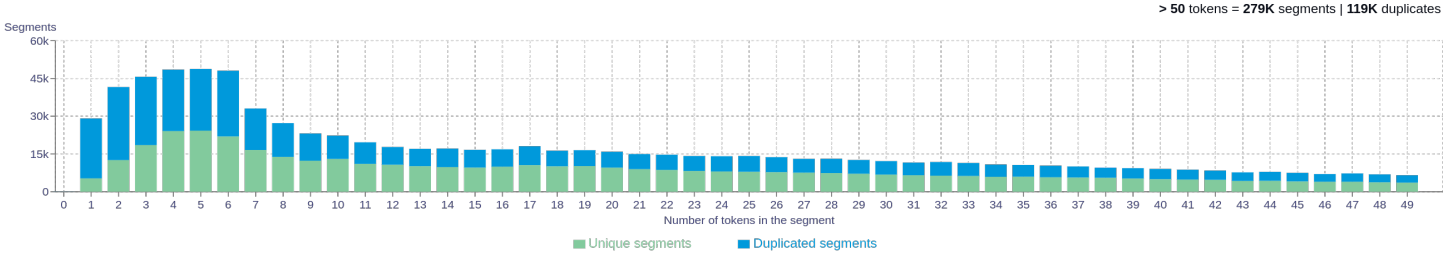
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (65K documents)

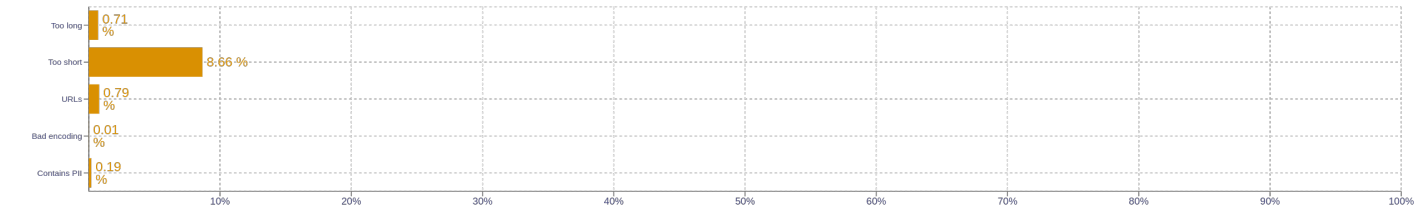


Segment length distribution by token

<= 49 tokens = **439K** segments | **410K** duplicates
> 50 tokens = **279K** segments | **119K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ኢርትራ | 192937 · ህዝቢ | 123530 · ኢዩ | 102095 · ከኣ | 101813 · ሰብ | 93382
2	ህዝቢ ኢርትራ | 34731 · ስርዓት ህግደፍ | 19331 · ቤት ፅሕፈት | 13319 · ቤተ ክርስቲያን | 12577 · ሕቡራት ሃገራት | 11653
3	ውድብ ሕቡራት ሃገራት | 4749 · ቅድስት ድንግል ማርያም | 2449 · ወወልድ ወመንፈስ ቅዱስ | 2071 · ሰልፊ ዲሞክራሲ ህዝቢ | 1713 · ሰምዖን ተ. ኣርኣያ | 1701
4	ሰልፊ ዲሞክራሲ ህዝቢ ኢርትራ | 1571 · ቤት ፅሕፈት ዜና ሰዲህኢ | 1434 · by ቤት ፅሕፈት ዜና | 1415 · ኣብ ወወልድ ወመንፈስ ቅዱስ | 1336 · በስመ ኣብ ወወልድ ወመንፈስ | 1318
5	by ቤት ፅሕፈት ዜና ሰዲህኢ | 1373 · በስመ ኣብ ወወልድ ወመንፈስ ቅዱስ | 1308 · Written by ቤት ፅሕፈት ዜና | 1109 · ኣብ ወወልድ ወመንፈስ ቅዱስ ኣሓዱ | 1040 · ወወልድ ወመንፈስ ቅዱስ ኣሓዱ ኣምላክ | 997

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