

General overview

Corpus	Date	Language
hplt-v3-tir_Ethi	9/18/2025	Tigrinya (ti)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
67,624	1,246,960	958,112 (76.84 %)	45M	190,686,473	452.41 MB

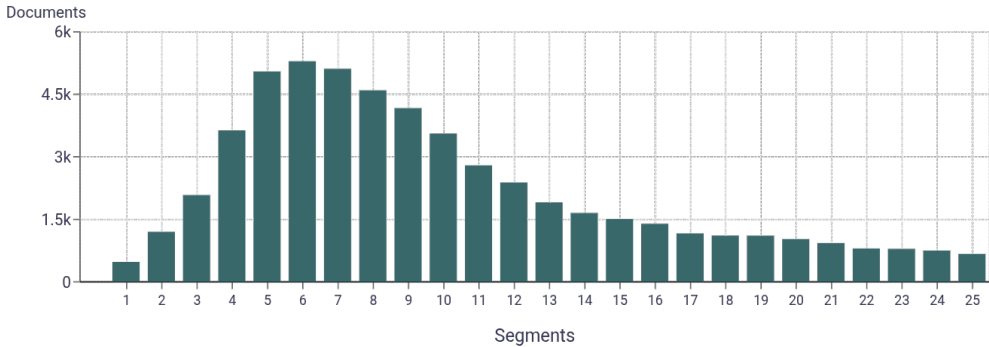
Top 10 domains

Domain	Docs	% of total
voanews.com	8.8K	12.97%
erena.org	5.3K	7.81%
assenna.com	4.4K	6.49%
vaticannews.va	3.9K	5.78%
harnnet.org	3.1K	4.57%
jw.org	2.2K	3.22%
bbc.com	1.9K	2.85%
harmonymedia.se	1.5K	2.21%
refugies.info	1.5K	2.16%
dendenmedia.com	1.5K	2.15%

Top 10 TLDs

Domain	Docs	% of total
com	31K	45.71%
org	19K	28.62%
va	5.3K	7.80%
se	3.6K	5.36%
net	1.7K	2.45%
info	1.6K	2.31%
de	753	1.11%
nl	623	0.92%
no	613	0.91%
ch	542	0.80%

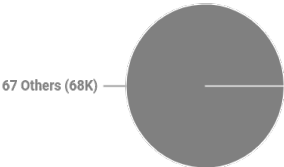
Documents size (in segments) ⓘ



≤ 25 segments **81.6%** (55K documents)  
> 25 segments **18.4%** (12K documents)

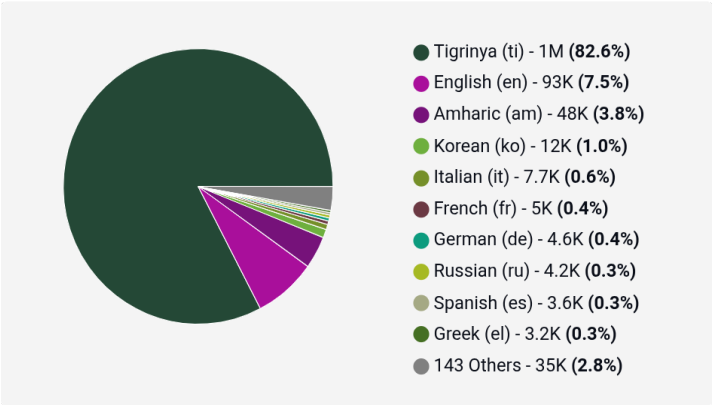
Document collections

CC = 91.01%  
IA = 8.99%

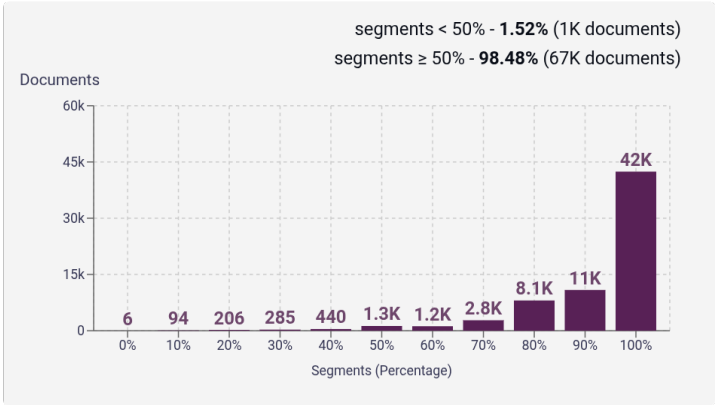


Language Distribution

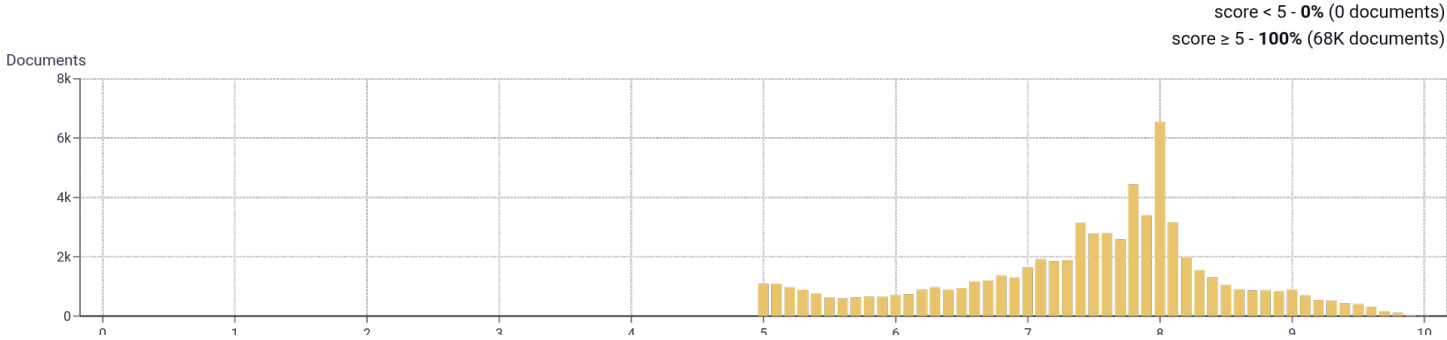
Number of segments in the Tigrinya (ti) corpus



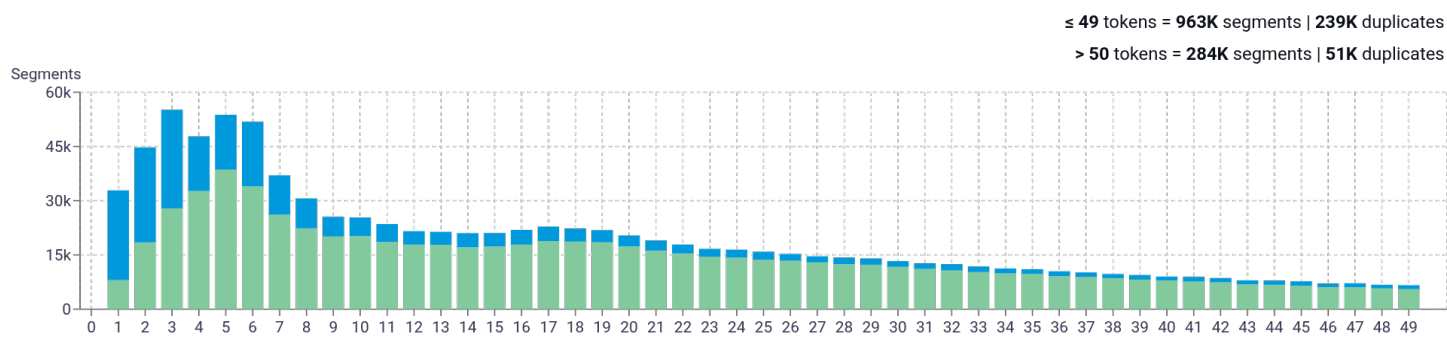
Percentage of segments in Tigrinya (ti) inside documents



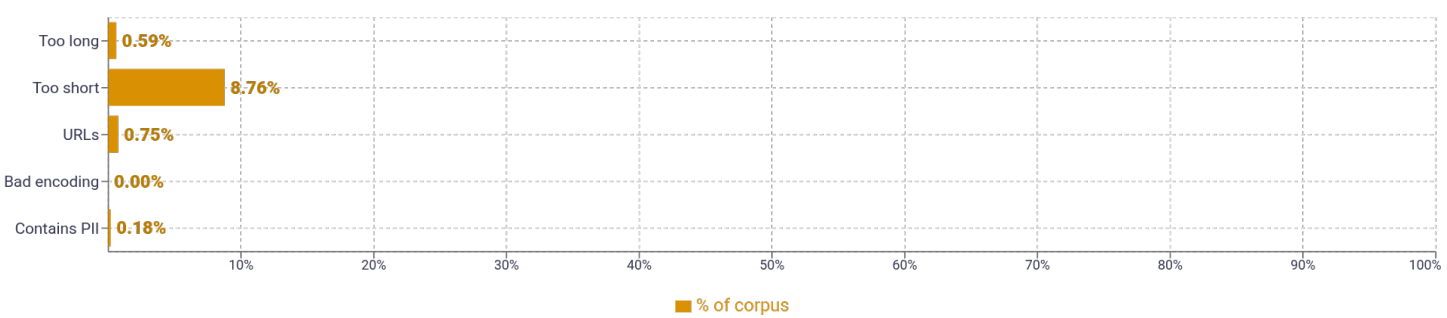
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ኢርትራ   176,623    ህዝቢ   121,978    ከአ   109,050    ሰብ   94,077    አዩ   79,628	
2	ህዝቢ ኢርትራ   34,190    ቤተ ክርስቲያን   16,019    ቤት ልሕዲት   13,647    ዓመተ ምሕረት.   11,837    ሕብረት ሃገራት   11,823	
3	ወድብ ሕብረት ሃገራት   4,063    ሰልፊ ዲሞክራሲ ህዝቢ   2,944    ወወልድ ወመንፈስ ቅዱስ   2,910    ዲሞክራሲ ህዝቢ ኢርትራ   2,846    ስምዖን ተ. አርአያ   2,646	
4	ሰልፊ ዲሞክራሲ ህዝቢ ኢርትራ   2,717    አብ ወወልድ ወመንፈስ ቅዱስ   2,225    በስመ አብ ወወልድ ወመንፈስ   2,163    ወወልድ ወመንፈስ ቅዱስ ኢሱዱ   2,006 ቅዱስ አቦና ር.ሊ.ዱ. ፍራንቸስኮ   1,933	
5	በስመ አብ ወወልድ ወመንፈስ ቅዱስ   2,144    አብ ወወልድ ወመንፈስ ቅዱስ ኢሱዱ   1,911    ወወልድ ወመንፈስ ቅዱስ ኢሱዱ አምላክ   1,769 ቅድስቲ መንበር ዚናን ከፍሊ ማሕተምን   1,241    written by ቤት ልሕዲት ዚና   1,082	

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				