

General overview

Corpus	Analytics 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	434.15 MB	180,568,832

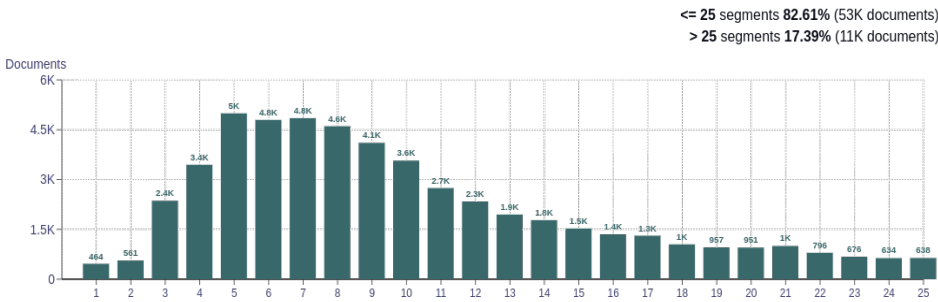
Top 10 domains

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

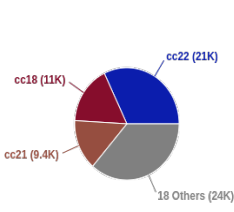
Top 10 TLDs

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)

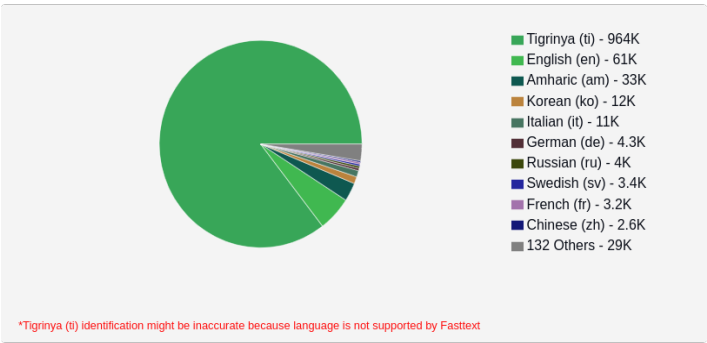


Documents by collection

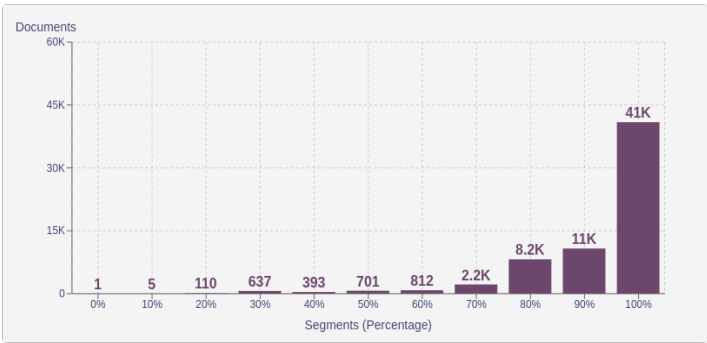


Language Distribution

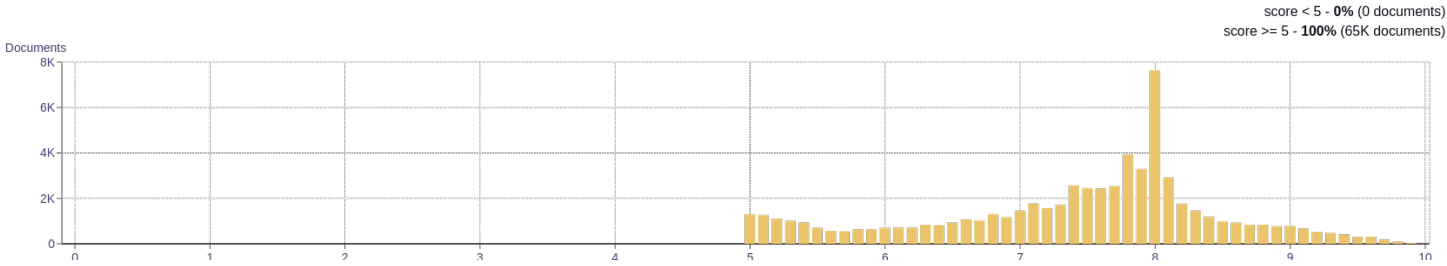
Number of segments



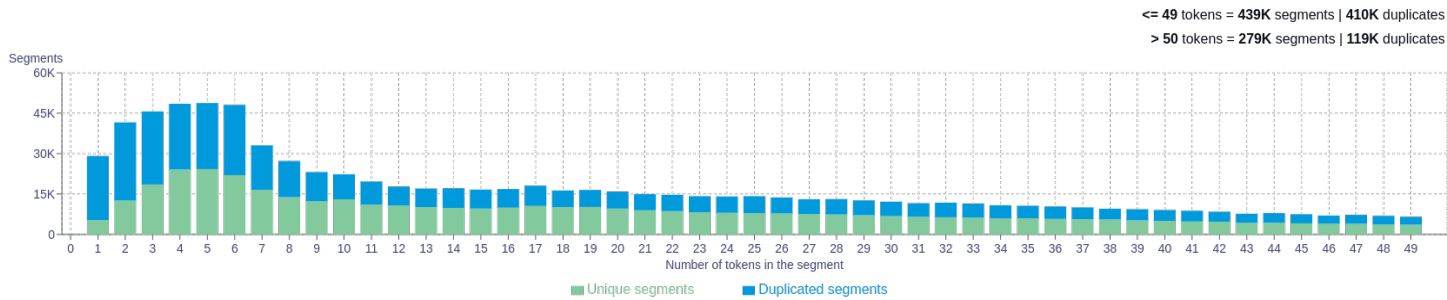
Percentage of segments in Tigrinya (ti) inside documents



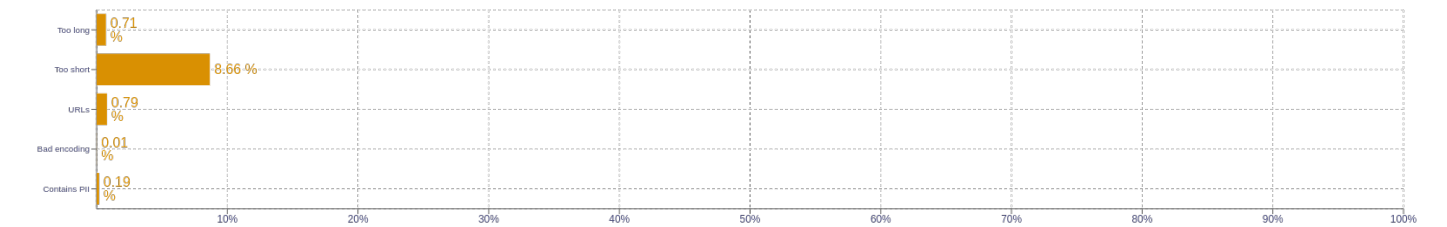
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ኢርትራ   192937</div> <div>ሀዘዚ   123530</div> <div>ኢዩ   102095</div> <div>የላ   101813</div> <div>ሱብ   93382</div>
2	<div>ሀዘዚ ኢርትራ   34731</div> <div>በርግት ሀገደሩ   19331</div> <div>ቤት ልክፊት   13319</div> <div>ቤት ዘርቡቴያን   12577</div> <div>ታላቅ ሃገራት   11653</div>
3	<div>ውድብ ታላቅ ሃገራት   4749</div> <div>ትድብት ድንግል ማርያም   2449</div> <div>መልድ መሪዎች ትድብ   2071</div> <div>ሰልፊ ሚኒስቴር ሀዘዚ   1713</div> <div>በምያን ተ. ኦርሲያ   1701</div>
4	<div>ሰልፊ ሚኒስቴር ሀዘዚ ኢርትራ   1571</div> <div>ቤት ልክፊት ዚና ሰላሳ   1434</div> <div>by ቤት ልክፊት ዚና   1415</div> <div>አብ መልድ መሪዎች ትድብ   1336</div> <div>በከፍ አብ መልድ መሪዎች   1318</div>
5	<div>by ቤት ልክፊት ዚና ሰላሳ   1373</div> <div>በከፍ አብ መልድ መሪዎች ትድብ   1308</div> <div>written by ቤት ልክፊት ዚና   1109</div> <div>አብ መልድ መሪዎች ትድብ አሁን   1040</div> <div>መልድ መሪዎች ትድብ አሁን እምላክ   997</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.stinfo.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>