HPLT Analytics report

General overview

Corpus	Date	Language
amh_Ethi.jsonl.tsv	9/6/2024	Amharic (am)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
295,542	7,005,835	3,859,756 (55.09 %)	226M	1,024,632,898	2.39 GB

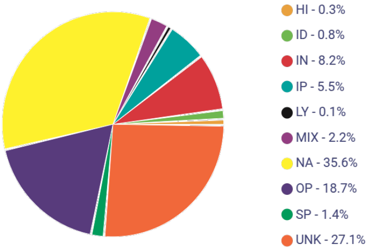
Top 10 domains

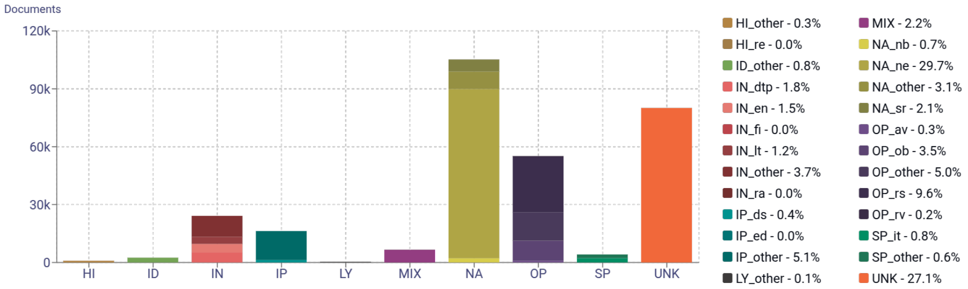
Domain	Docs	% of total
ethiopianreport...	15K	4.92%
wordpress.com	14K	4.88%
voanews.com	12K	4.15%
addisadmassnews...	11K	3.84%
ethioreference.com	8.8K	2.99%
blogspot.com	7.1K	2.40%
blogspot.no	7K	2.35%
goolgule.com	6.2K	2.10%
wikipedia.org	5.9K	1.98%
dw.com	5.6K	1.89%

Top 10 TLDs

Domain	Docs	% of total
com	203K	68.65%
org	33K	11.11%
net	8.5K	2.86%
no	7.4K	2.49%
gov.et	5.3K	1.81%
et	5.3K	1.79%
news	2.6K	0.87%
info	2.4K	0.80%
ch	2.2K	0.73%
us	1.9K	0.65%

Register labels

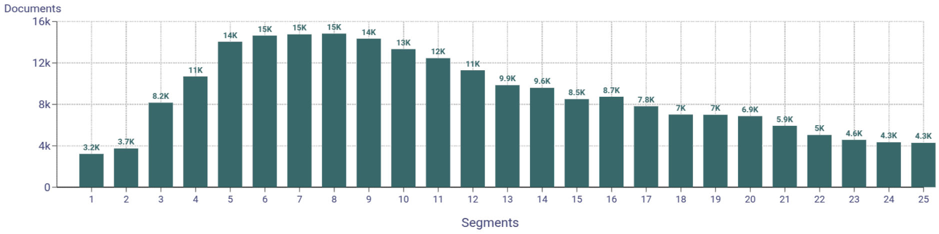




🤖 **MT**:17.9% | 53K Documents

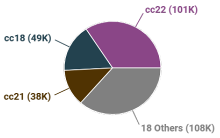
Documents size (in segments)

<= 25 segments **76.14%** (225K documents)
> 25 segments **23.86%** (71K documents)



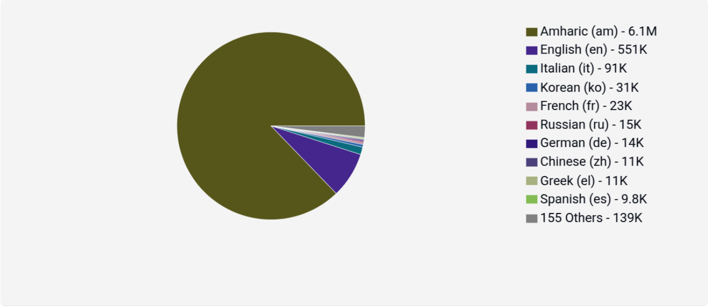
Documents by collection

CC = 70.50%
IA = 29.50%

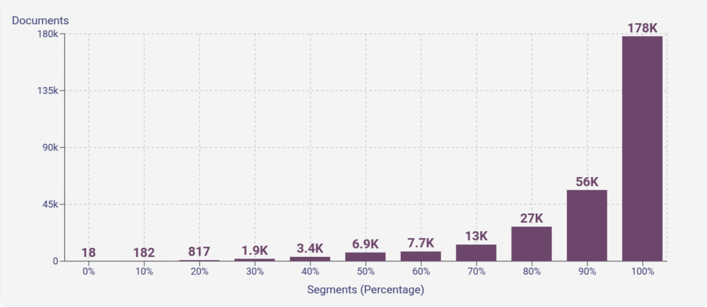


Language Distribution

Number of segments in the Amharic (am) corpus

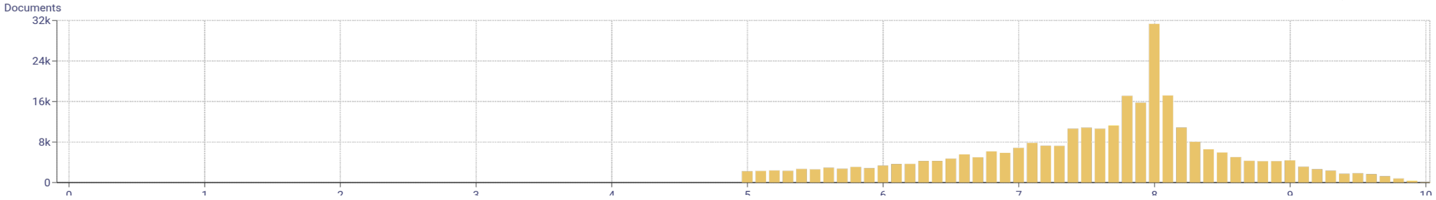


Percentage of segments in Amharic (am) inside documents



Distribution of documents by document score

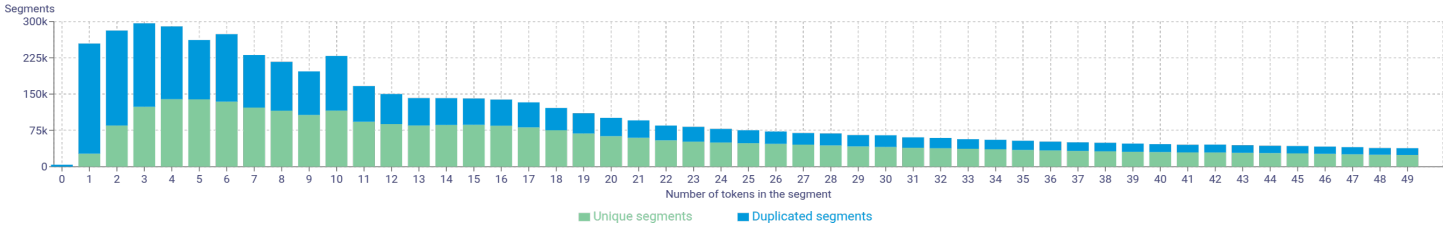
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (296K documents)



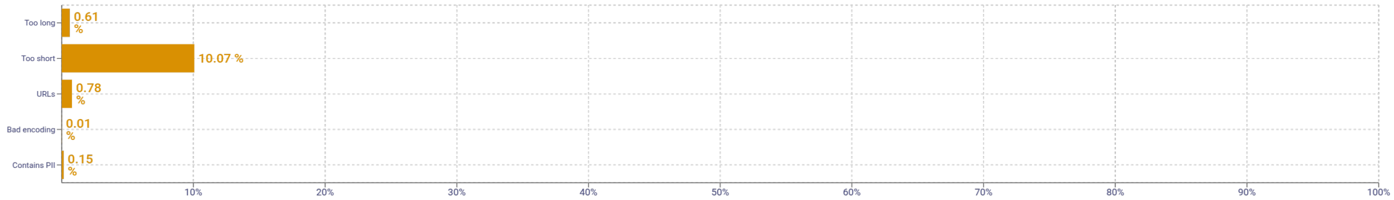
Segment length distribution by token

≤ 49 tokens = **3M** segments | **2.6M** duplicates

> 50 tokens = **1.5M** segments | **589K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ቤት | 399734 ሰው | 351453 ሰዎች | 288104 the | 268082 ብዙ | 264121
2	ዓመተ ምህረት | 131922 ዓመተ ምህረት. | 84621 አዲስ አበባ | 70843 ምክር ቤት | 48490 ፍርድ ቤት | 47456
3	አዳዲስ ግምገማዎች በይፋ | 11644 ተወካዮች ምክር ቤት | 9813 የኢትዮጵያ ኦርቶዶክስ ተዋሕዶ | 9769 ተዋሕዶ ቤተ ክርስቲያን | 9612 leave a comment | 8551
4	ኦርቶዶክስ ተዋሕዶ ቤተ ክርስቲያን | 8777 size increase font size | 7105 size decrease font size | 7105 font size increase font | 7105 font size decrease font | 7105
5	size decrease font size increase | 7105 font size increase font size | 7105 font size decrease font size | 7105 decrease font size increase font | 7105 የኢትዮጵያ ኦርቶዶክስ ተዋሕዶ ቤተ ክርስቲያን | 5385

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