

General overview

Corpus	Date	Language
hplt-v3-amh_Ethi	9/16/2025	Amharic

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
571,240	12,248,888	9,343,853 (76.28 %)	367M	1,653,669,891	3.86 GB

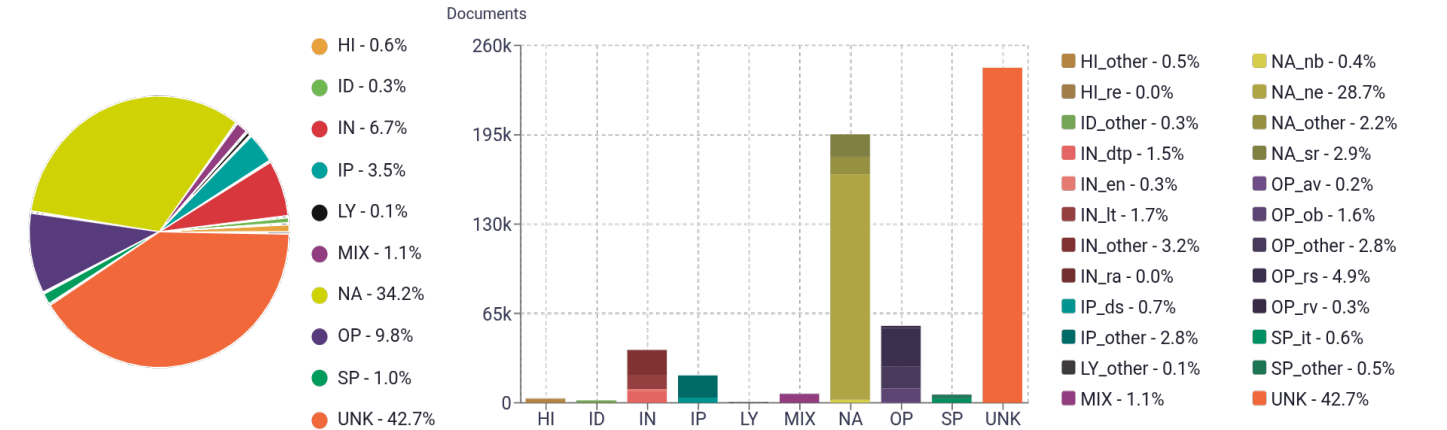
Top 10 domains

Domain	Docs	% of total
ethiopianreport...	19K	3.39%
addisadmassnews...	18K	3.23%
fanabc.com	18K	3.20%
soccerethiopia.net	15K	2.68%
wordpress.com	14K	2.45%
dw.com	11K	1.87%
voanews.com	10K	1.79%
al-ain.com	9.7K	1.70%
ethioreference.com	9.7K	1.70%
ethioexplorer.com	9.7K	1.69%

Top 10 TLDs

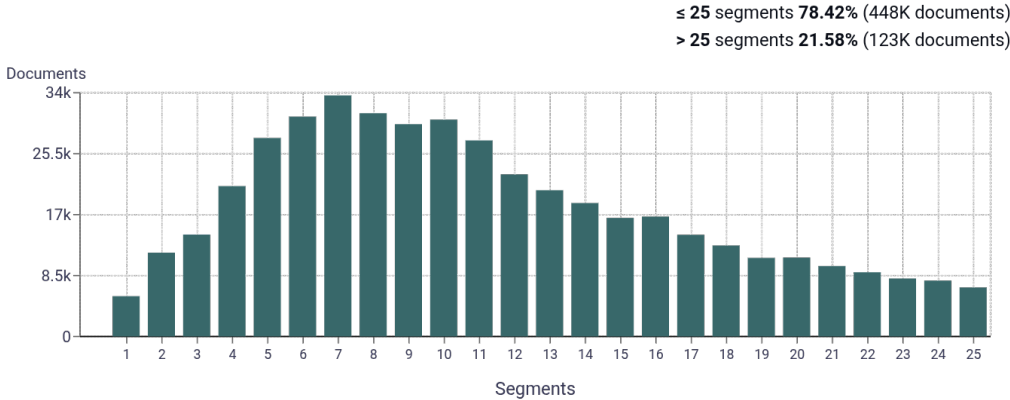
Domain	Docs	% of total
com	419K	73.37%
org	37K	6.49%
net	23K	4.00%
gov.et	13K	2.22%
et	12K	2.16%
zone	7.6K	1.32%
news	4.6K	0.80%
no	3.9K	0.68%
edu.et	3.6K	0.63%
us	3.3K	0.58%

Register labels

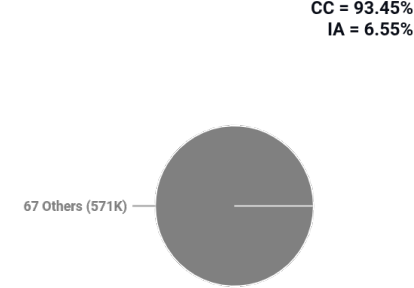


MT:36.7% | 210K Documents

Documents size (in segments) ⓘ

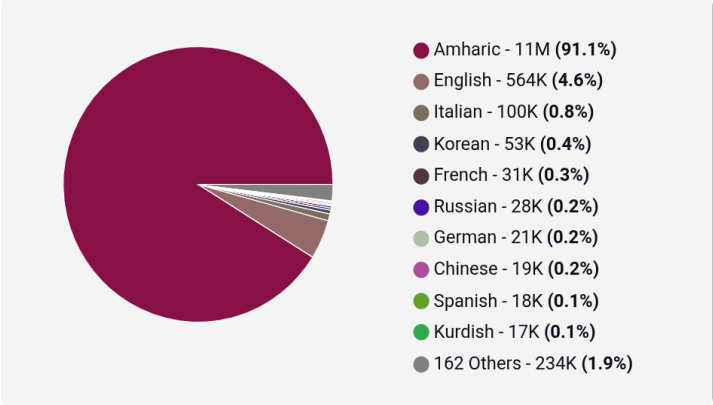


Document collections

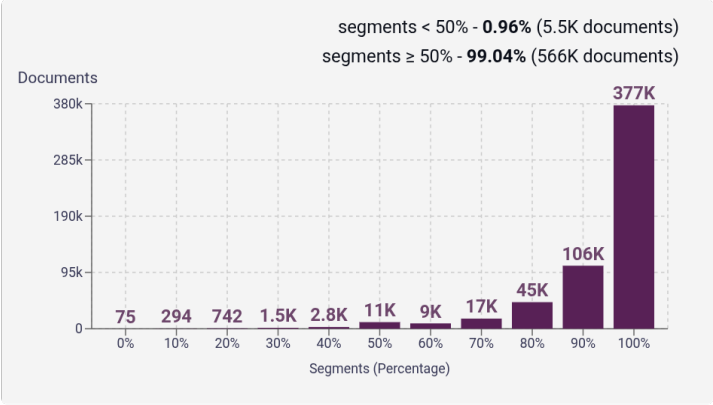


Language Distribution

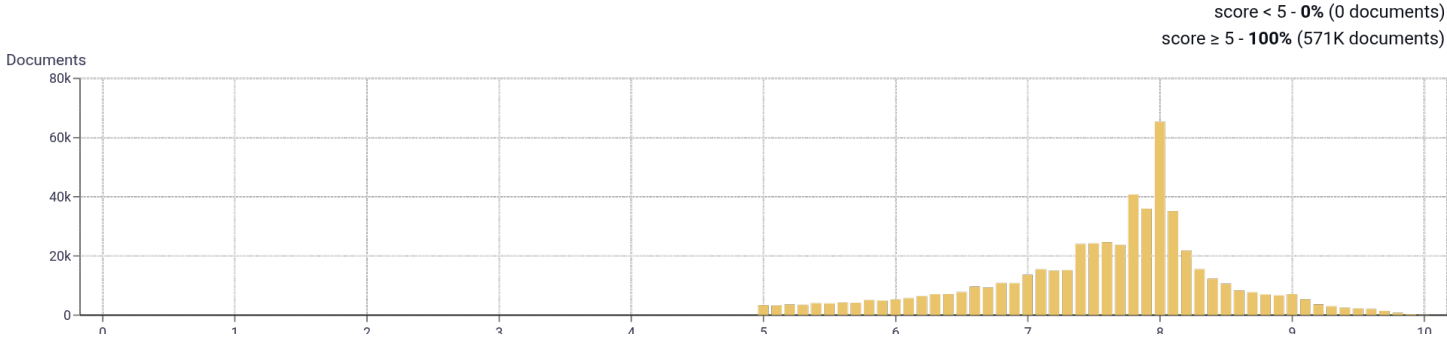
Number of segments in the Amharic corpus



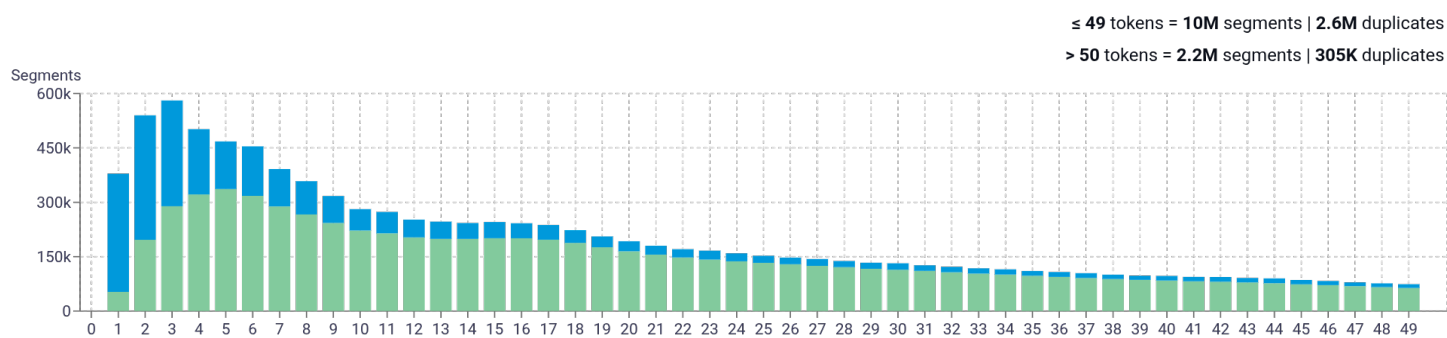
Percentage of segments in Amharic inside documents



Distribution of documents by document score

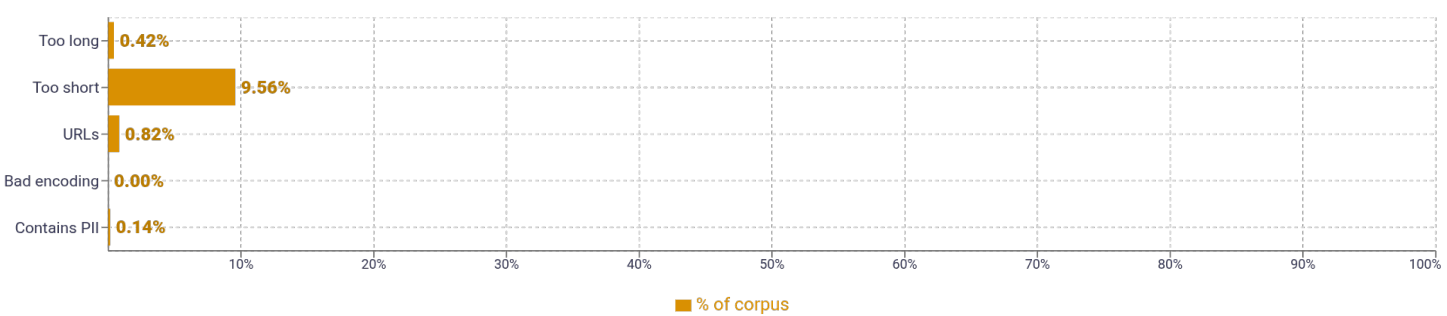


Segment length distribution by token



≤ 49 tokens = 10M segments | 2.6M duplicates  
> 50 tokens = 2.2M segments | 305K duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>ቤት   588,029</div> <div>ሰው   526,160</div> <div>ብዙ   515,527</div> <div>ሰዎች   475,288</div> <div>ከፍተኛ   471,026</div>	
2	<div>ዓመተ ምህረት   166,325</div> <div>ትድሚያ የሚሰጡዋቸውን   110,154</div> <div>እዲስ አበባ   109,711</div> <div>ዓመተ ምህረት.   100,547</div> <div>ሙሉ በሙሉ   90,703</div>	
3	<div>አዳዲስ ግምገማዎች በይፋ   28,412</div> <div>ኢፍ.ቢ.ሲ.   17,971</div> <div>ሸቀጣ ሸቀጣ ሸቀጣ   17,619</div> <div>ተወካዮች ምክር ቤት   16,252</div> <div>በላጀመሩ ትድሚያ የሚሰጡዋቸውን   15,924</div>	
4	<div>ሸቀጣ ሸቀጣ ሸቀጣ ሸቀጣ   16,861</div> <div>እስተያየት ለመስጠት የመጀመሪያው ደውኦ   12,579</div> <div>ብት ብት ብት ብት   11,189</div> <div>ግምገማዎች ላይ ያለው ስምዎ   10,205</div> <div>የቅርብ ጊዜውን መልስክት ይመልከቱ   8,428</div>	
5	<div>ሸቀጣ ሸቀጣ ሸቀጣ ሸቀጣ ሸቀጣ   16,139</div> <div>ብት ብት ብት ብት ብት   10,990</div> <div>በቀዳሚ ግምገማዎች ላይ ያለው ስምዎ   7,038</div> <div>ኢ.ኢ..ኢ. ኢ.ኢ..ኢ. ኢ.ኢ..ኢ. ኢ.ኢ..ኢ. ኢ.ኢ..ኢ.   6,085</div> <div>ቶሎ ቶሎ ቶሎ ቶሎ ቶሎ   4,742</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>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
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				