HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-som_Latn	9/18/2025	Somali

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,423,820	18,734,302	14,006,522 (74.76 %)	534M	3,142,032,985	2.94 GB

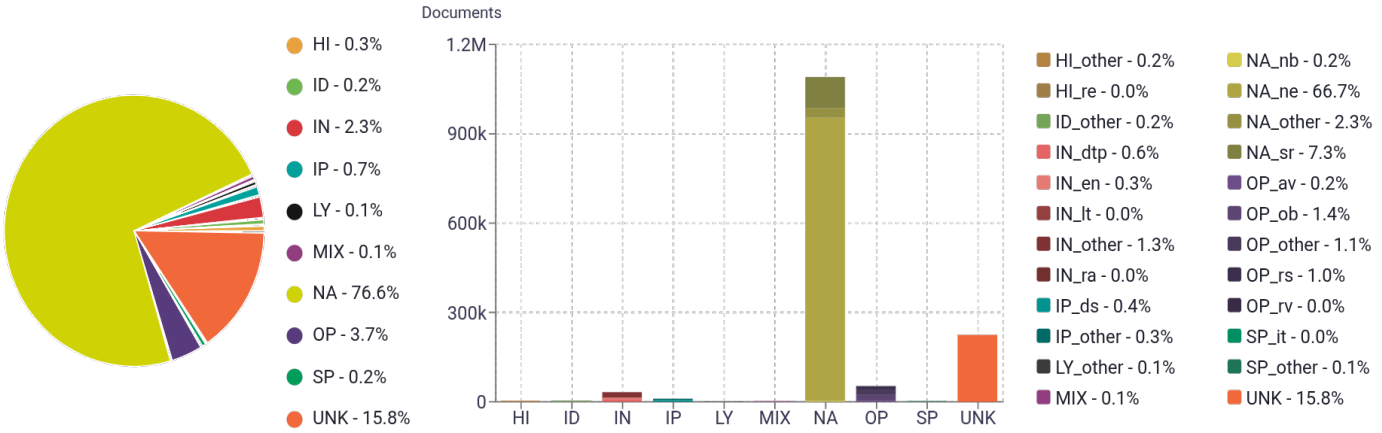
Top 10 domains

Domain	Docs	% of total
caasimada.net	35K	2.45%
goobjoog.com	32K	2.24%
radiodalsan.com	22K	1.53%
radiomuqdisho.net	21K	1.49%
goolfm.net	21K	1.45%
voasomali.com	17K	1.16%
wordpress.com	14K	0.99%
horseedmedia.net	14K	0.97%
starfm.co.ke	13K	0.92%
somaliweyn.org	12K	0.86%

Top 10 TLDs

Domain	Docs	% of total
com	942K	66.19%
net	297K	20.84%
org	58K	4.09%
so	34K	2.37%
co.ke	14K	0.98%
ca	9.8K	0.69%
online	9.6K	0.68%
se	7.4K	0.52%
info	5.9K	0.41%
co.uk	4.4K	0.31%

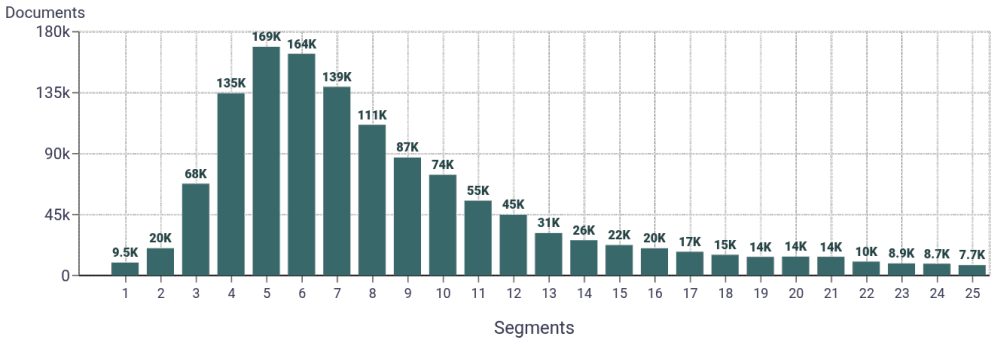
Register labels



MT:11.3% | 161K Documents

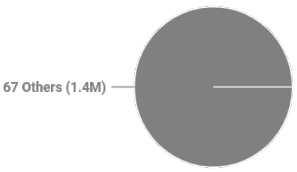
Documents size (in segments)

≤ 25 segments **90.28%** (1.3M documents)
> 25 segments **9.72%** (138K documents)



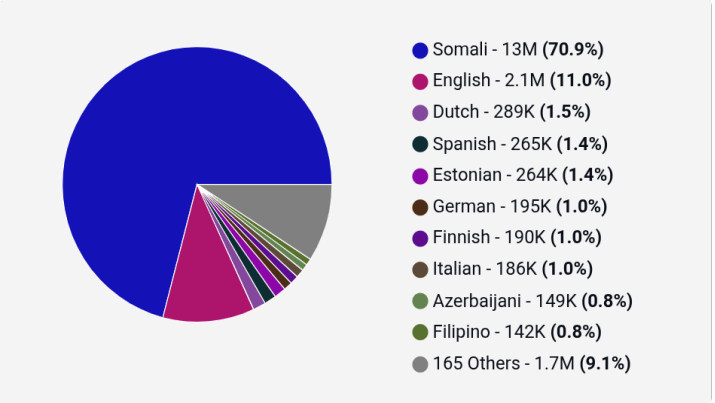
Document collections

CC = 82.69%
IA = 17.31%

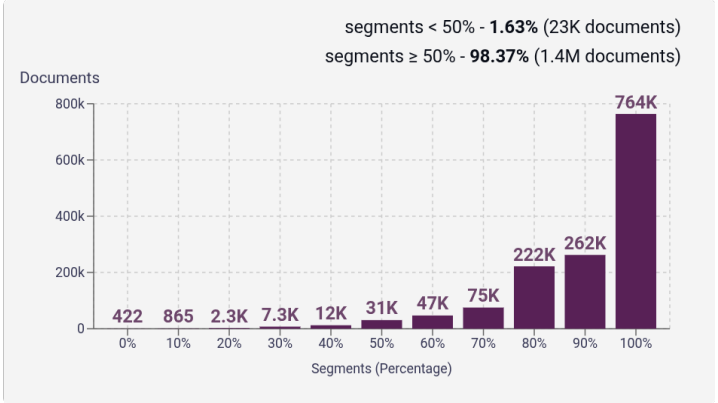


Language Distribution

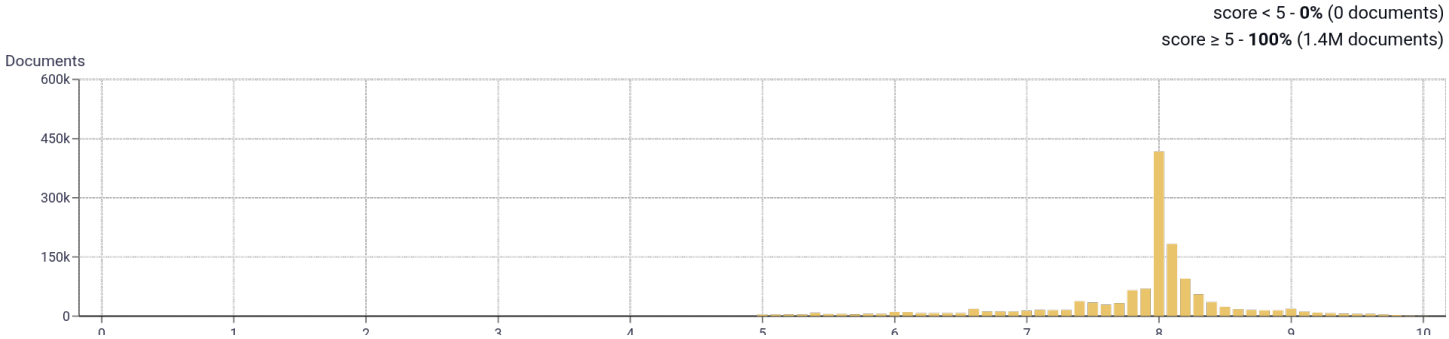
Number of segments in the Somali corpus



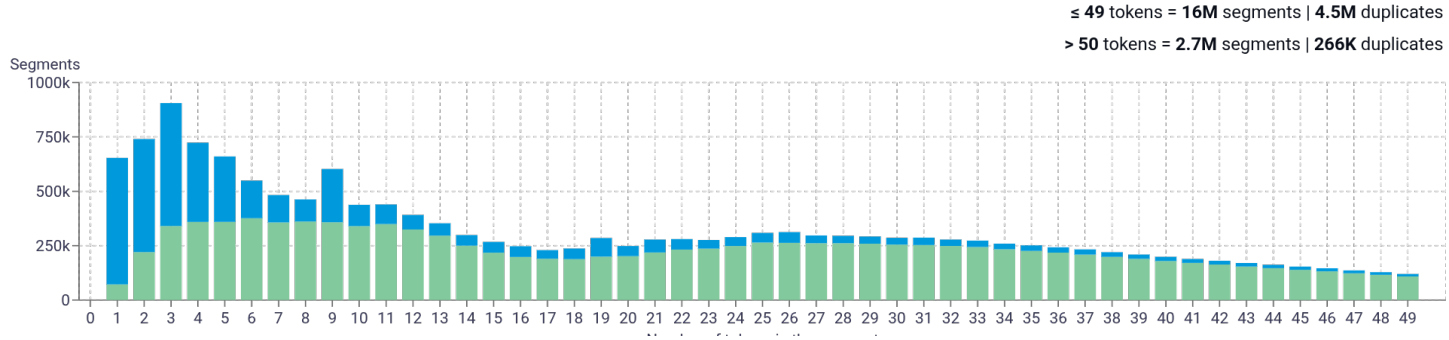
Percentage of segments in Somali inside documents



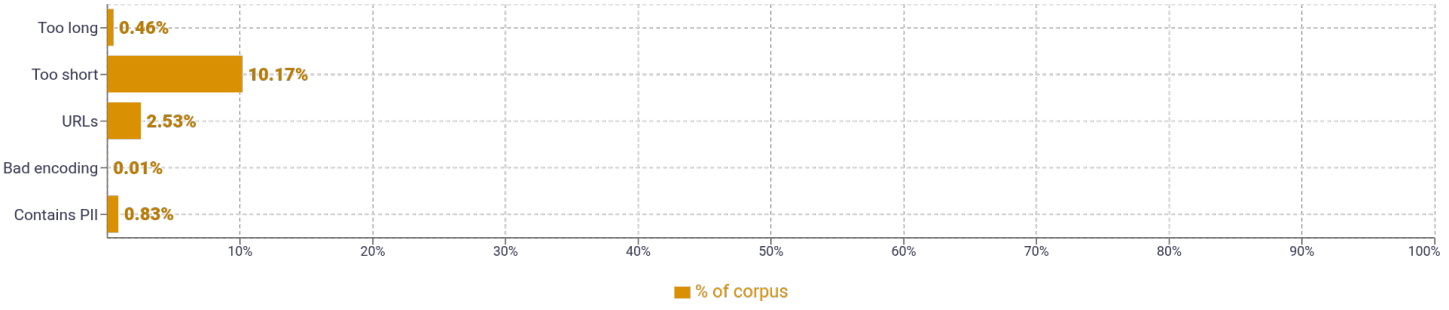
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ah | 8,703,434 · u | 6,814,235 · lagu | 2,652,091 · waxaa | 2,541,279 · aan | 2,054,425
2	mid ah | 774,207 · kala duwan | 373,842 · magaalada muqdisho | 245,188 · isla markaana | 242,023 · gaar ah | 236,340
3	mid ka mid | 143,846 · qaar ka mid | 123,521 · waxaa ka mid | 75,913 · wax soo saarka | 66,243 · u baahan tahay | 60,168
4	mid ka mid ah | 137,364 · qaar ka mid ah | 118,839 · waxaa ka mid ah | 58,693 · ah oo ku saabsan | 29,470 · wax ku ool ah | 24,366
5	be the first to comment | 18,208 · badan oo ka mid ah | 16,738 · kuwa ugu horreeya ee faallo | 12,248 · madaxweynaha jamhuuriyadda federaalka soomaaliya mudane | 11,092 · sida uu hadalka u dhigay | 10,615

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.	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				