HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-swh_Latn	9/24/2025	Swahili

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,935,907	43,154,287	27,265,037 (63.18 %)	1.1B	6,168,365,814	5.77 GB

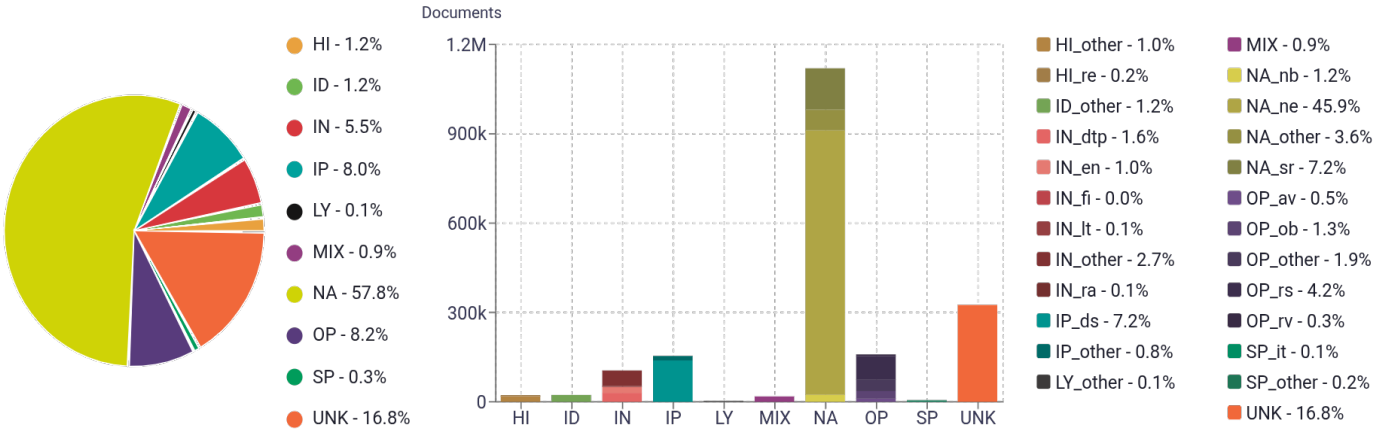
Top 10 domains

Domain	Docs	% of total
blogspot.com	148K	7.66%
airbnb.com	116K	5.97%
tuko.co.ke	56K	2.89%
nation.co.ke	31K	1.63%
mtanzania.co.tz	31K	1.61%
dw.com	30K	1.56%
millardayo.com	25K	1.27%
zanzinews.com	24K	1.22%
un.org	21K	1.08%
habarileo.co.tz	20K	1.04%

Top 10 TLDs

Domain	Docs	% of total
com	1.1M	54.38%
co.tz	296K	15.28%
co.ke	142K	7.33%
org	141K	7.28%
go.tz	83K	4.29%
net	32K	1.65%
fr	22K	1.12%
cn	14K	0.72%
info	11K	0.56%
va	11K	0.56%

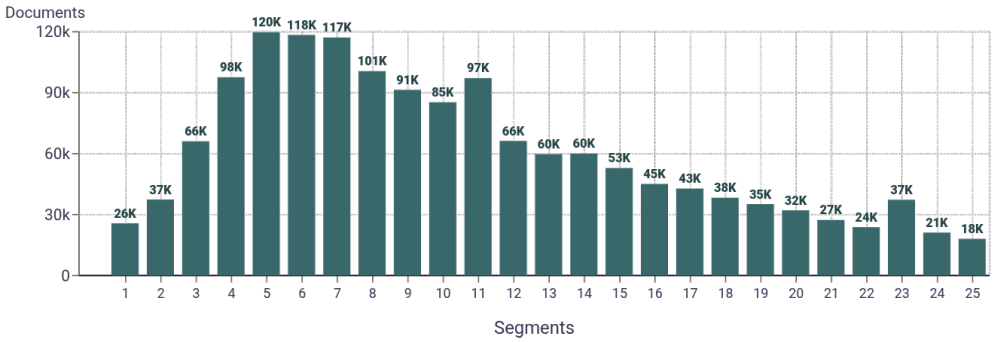
Register labels



MT:17.1% | 332K Documents

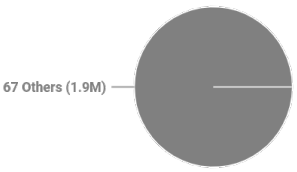
Documents size (in segments)

≤ 25 segments **78.28%** (1.5M documents)
> 25 segments **21.72%** (420K documents)



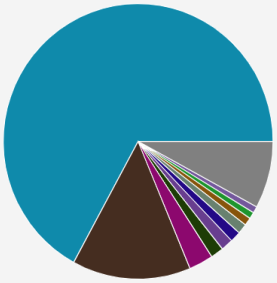
Document collections

CC = 87.25%
IA = 12.75%



Language Distribution

Number of segments in the Swahili corpus

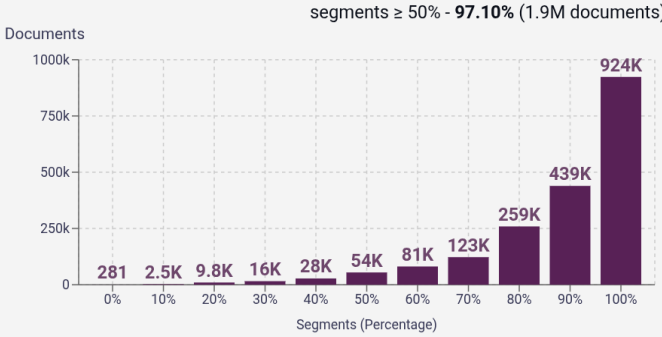


- Swahili - 29M **(67.1%)**
- English - 6.1M **(14.2%)**
- Italian - 1.3M **(2.9%)**
- Indonesian - 669K **(1.6%)**
- Spanish - 616K **(1.4%)**
- Esperanto - 551K **(1.3%)**
- German - 508K **(1.2%)**
- Dutch - 402K **(0.9%)**
- Finnish - 337K **(0.8%)**
- Polish - 329K **(0.8%)**
- 165 Others - 3.4M **(7.9%)**

*Swahili identification might be inaccurate because it is not supported by FastSpell.

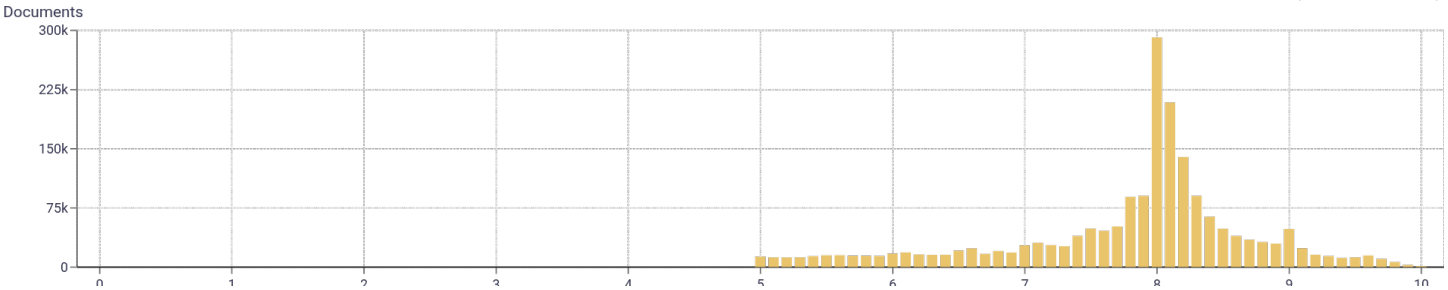
Percentage of segments in Swahili inside documents

segments < 50% - **2.90%** (56K documents)
segments ≥ 50% - **97.10%** (1.9M documents)



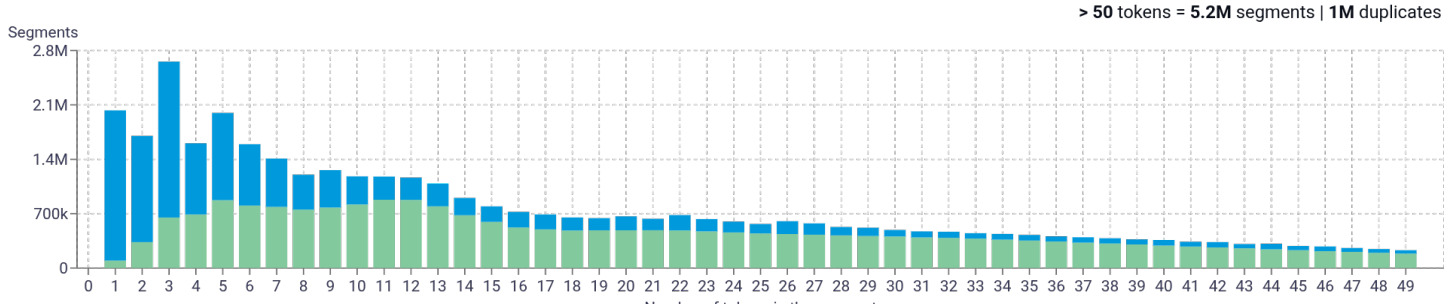
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.9M documents)

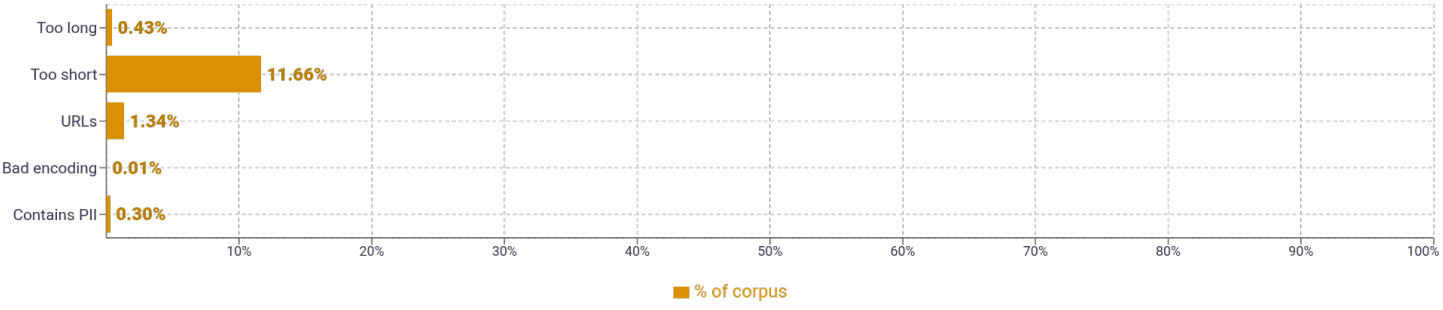


Segment length distribution by token

≤ 49 tokens = **38M** segments | **15M** duplicates
> 50 tokens = **5.2M** segments | **1M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	vya | 2,917,481 · hii | 2,650,121 · kazi | 2,097,666 · kati | 2,054,972 · ambayo | 2,004,131
2	dar es | 580,604 · es salaam | 567,271 · mwanzo huko | 329,277 · mwaka huu | 304,610 · kufanya kazi | 275,096
3	ukadiriaji wa wastani | 1,006,765 · dar es salaam | 561,846 · ukurasa wa mwanzo | 332,022 · post a comment | 242,791 · jijini dar es | 218,564
4	ukurasa wa mwanzo huko | 329,225 · jijini dar es salaam | 210,997 · nyumba ya kupangisha huko | 147,945 · kukodisha wakati wa likizo | 88,378 · pata na uweke nafasi | 64,771
5	jamhuri ya muungano wa tanzania | 141,688 · rais wa jamhuri ya muungano | 84,749 · pata nafasi ambayo ni sahihi | 63,166 · nafasi ambayo ni sahihi kwako | 63,166 · mwenyekiti wa baraza la mapinduzi | 50,189

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				