HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lug_Latn	9/18/2025	Ganda

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
49,599	738,143	619,748 (83.96 %)	22M	123,290,120	120.37 MB

Top 10 domains

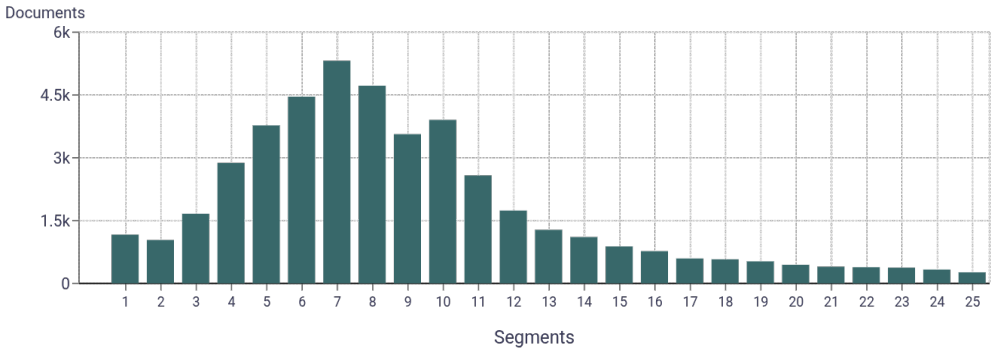
Domain	Docs	% of total
bukedde.co.ug	18K	36.24%
dembefm.ug	6.3K	12.73%
cbsfm.ug	3.7K	7.37%
radiosimba.ug	2.5K	4.95%
gambuuze.ug	2.1K	4.14%
galaxyfm.co.ug	1.3K	2.56%
wordproject.org	1.2K	2.34%
watchdoguganda.com	1K	2.11%
ssekanolya.co.ug	1K	2.08%
jw.org	1K	2.05%

Top 10 TLDs

Domain	Docs	% of total
co.ug	22K	43.38%
ug	16K	31.53%
com	7.2K	14.52%
org	3.5K	7.06%
is	788	1.59%
net	262	0.53%
info	118	0.24%
or.ug	99	0.20%
sv	72	0.15%
eu	51	0.10%

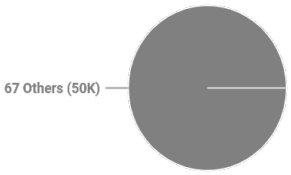
Documents size (in segments)

≤ 25 segments **90.24%** (45K documents)
> 25 segments **9.76%** (4.8K documents)



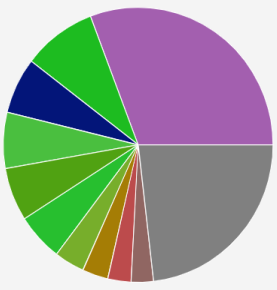
Document collections

CC = 91.83%
IA = 8.17%



Language Distribution

Number of segments in the Ganda corpus

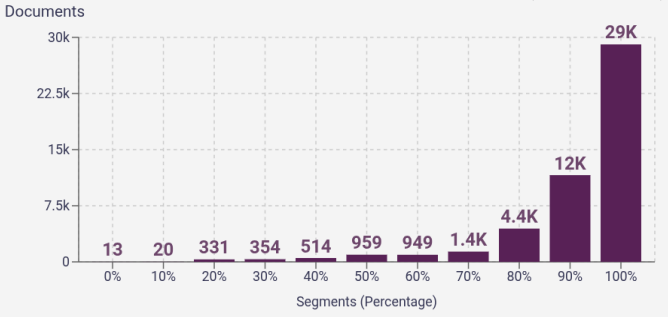


*Ganda identification might be inaccurate because it is not supported by FastSpell.

- English - 227K **(30.8%)**
- Waray - 65K **(8.8%)**
- Filipino - 49K **(6.6%)**
- Swahili - 49K **(6.6%)**
- Finnish - 46K **(6.3%)**
- Esperanto - 41K **(5.6%)**
- Indonesian - 27K **(3.6%)**
- Spanish - 23K **(3.1%)**
- Polish - 21K **(2.8%)**
- Sundanese - 20K **(2.7%)**
- 150 Others - 171K **(23.2%)**

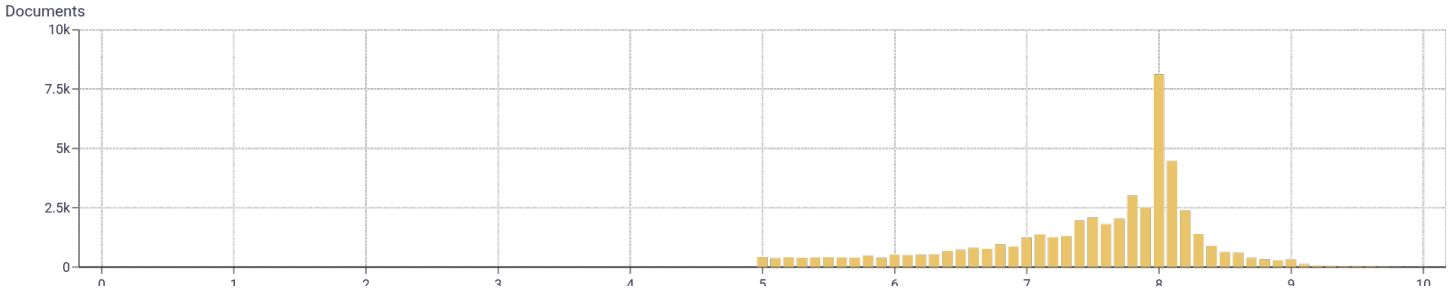
Percentage of segments in Ganda inside documents

segments < 50% - **2.48%** (1.2K documents)
segments ≥ 50% - **97.52%** (48K documents)



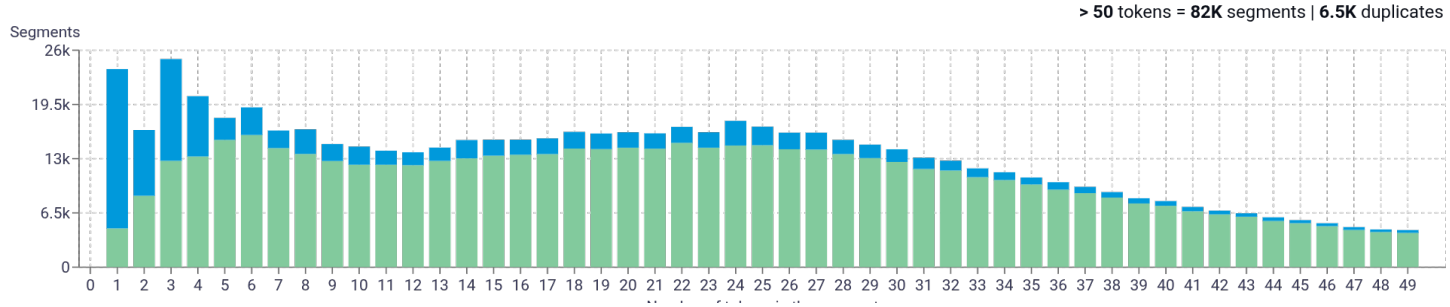
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (50K documents)

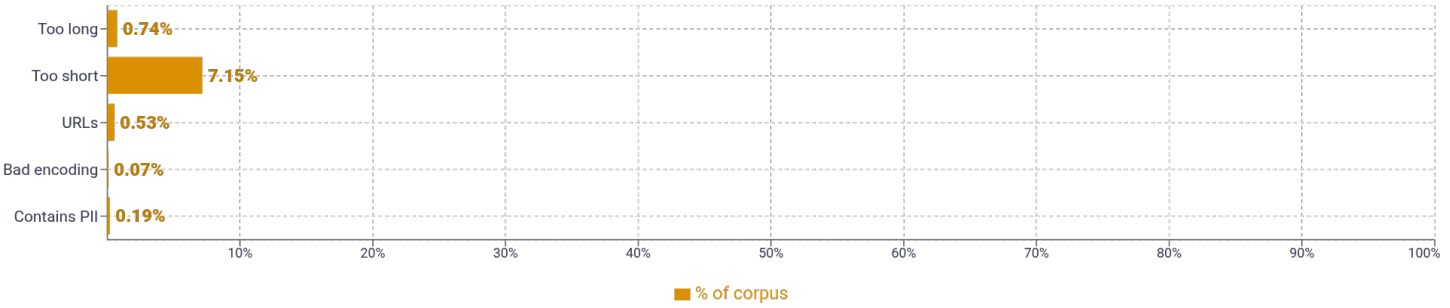


Segment length distribution by token

≤ 49 tokens = **656K** segments | **112K** duplicates
> 50 tokens = **82K** segments | **6.5K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	n | 443,023 · ng | 123,746 · abantu | 77,794 · y | 74,283 · bw | 66,931
2	mukama katonda | 7,264 · wamu n | 6,479 · disitulikiti y | 4,325 · poliisi y | 3,574 · mukama n | 3,517
3	abaana ba isiraeri | 2,286 · omwogezi wa poliisi | 2,063 · charles peter mayiga | 1,984 · maaso ga mukama | 1,794 · mukama katonda wo | 1,521
4	endagaano enkadde n'endagaano empya | 1,215 · do you have a | 1,063 · you have a story | 1,062 · or an opinion to | 1,062 · to share with us | 1,061
5	do you have a story | 1,062 · or an opinion to share | 1,061 · your community or an opinion | 1,060 · you have a story in | 1,060 · story in your community or | 1,060

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				