HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-run_Latn	9/18/2025	Rundi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
235,311	2,868,404	2,236,602 (77.97 %)	88M	492,277,507	478.44 MB

Top 10 domains

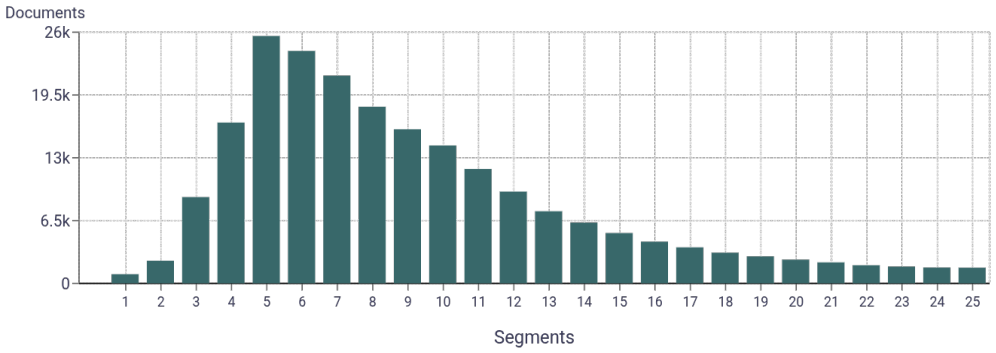
Domain	Docs	% of total
radiyoyacuvoa.com	20K	8.40%
igihe.bi	16K	7.01%
igihe.com	14K	5.79%
umuryango.rw	12K	5.00%
kigalitoday.com	9.8K	4.18%
bbc.com	8.2K	3.49%
inyarwanda.com	7.7K	3.26%
indundi.com	7.4K	3.16%
taarifa.rw	6.2K	2.66%
yegob.rw	5.6K	2.39%

Top 10 TLDs

Domain	Docs	% of total
com	133K	56.37%
rw	60K	25.29%
bi	18K	7.65%
org	6.8K	2.90%
co.rw	6.8K	2.90%
net	5.4K	2.27%
fr	2.6K	1.10%
info	492	0.21%
gov.rw	473	0.20%
tv	374	0.16%

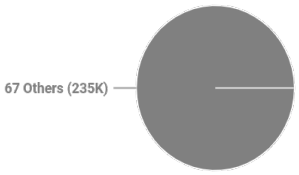
Documents size (in segments)

≤ 25 segments **91.16%** (215K documents)
> 25 segments **8.84%** (21K documents)



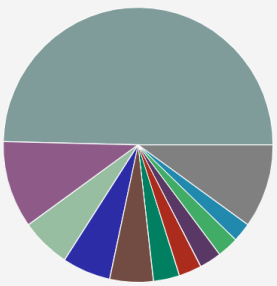
Document collections

CC = 89.94%
IA = 10.06%



Language Distribution

Number of segments in the Rundi corpus

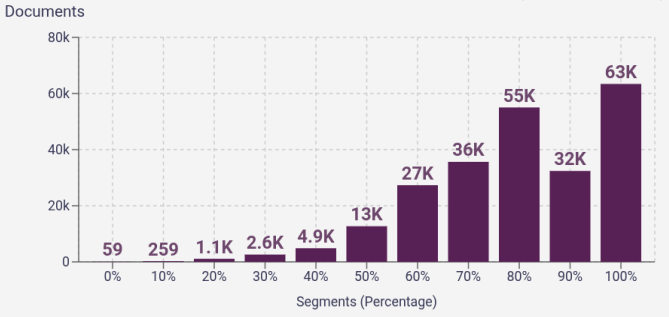


*Rundi identification might be inaccurate because it is not supported by FastSpell.

- English - 1.4M **(49.6%)**
- Swahili - 294K **(10.3%)**
- Esperanto - 170K **(5.9%)**
- Filipino - 166K **(5.8%)**
- Indonesian - 150K **(5.2%)**
- French - 89K **(3.1%)**
- German - 79K **(2.7%)**
- Spanish - 75K **(2.6%)**
- Turkish - 73K **(2.5%)**
- Polish - 65K **(2.3%)**
- 153 Others - 283K **(9.9%)**

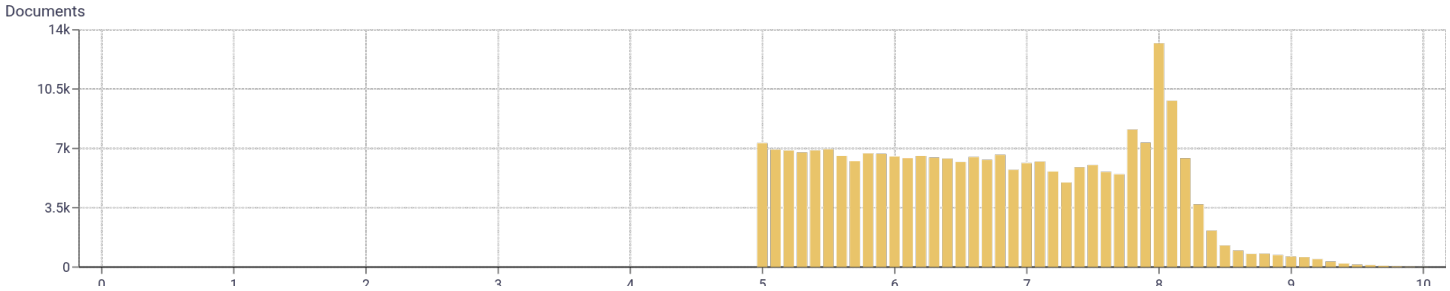
Percentage of segments in Rundi inside documents

segments < 50% - **3.77%** (8.9K documents)
segments ≥ 50% - **96.23%** (226K documents)



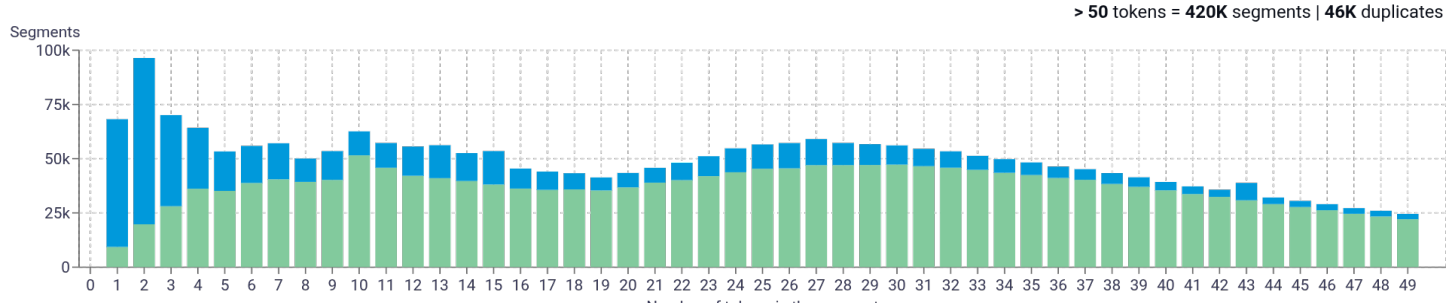
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (235K documents)

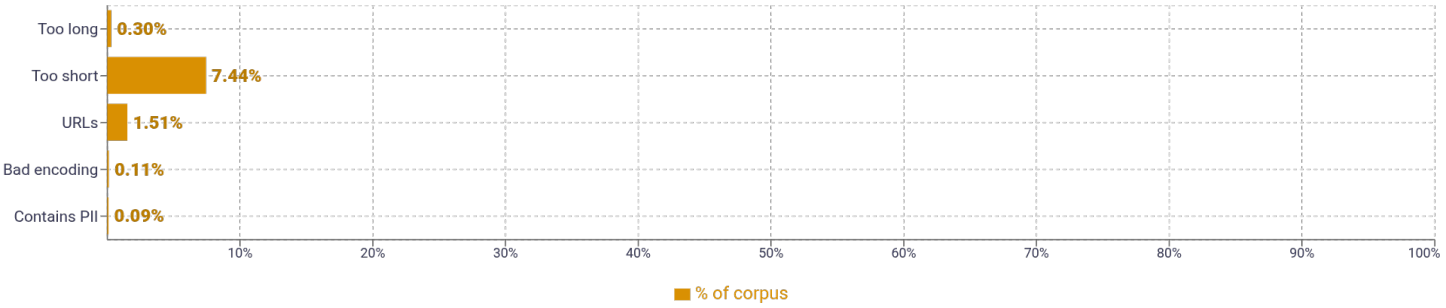


Segment length distribution by token

≤ 49 tokens = **2.4M** segments | **592K** duplicates
> 50 tokens = **420K** segments | **46K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	n | 1,312,664 · y | 905,889 · w | 594,835 · rwanda | 378,702 · u | 271,141
2	u rwanda | 199,919 · nyuma y | 87,796 · rayon sports | 86,366 · umukuru w | 62,913 · apr fc | 60,987
3	leta zunze ubumwe | 25,863 · ubumwe za amerika | 20,163 · jenoside yakorewe abatutsi | 17,640 · ikipe ya rayon | 13,959 · demokarasi ya congo | 13,083
4	zunze ubumwe za amerika | 19,907 · ikipe ya rayon sports | 11,863 · iharanira demokarasi ya congo | 11,146 · leta zunze ubumwe z | 9,152 · ikipe ya apr fc | 8,305
5	leta zunze ubumwe za amerika | 15,021 · repubulika iharanira demokarasi ya congo | 10,307 · igitekerezo cyawe kigaragara nyuma y | 4,903 · igitekerezo cyanyu gishobora kutagaragara hano | 4,903 · ibi bidakurikijwe igitekerezo cyanyu gishobora | 4,903

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				