

General overview

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

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

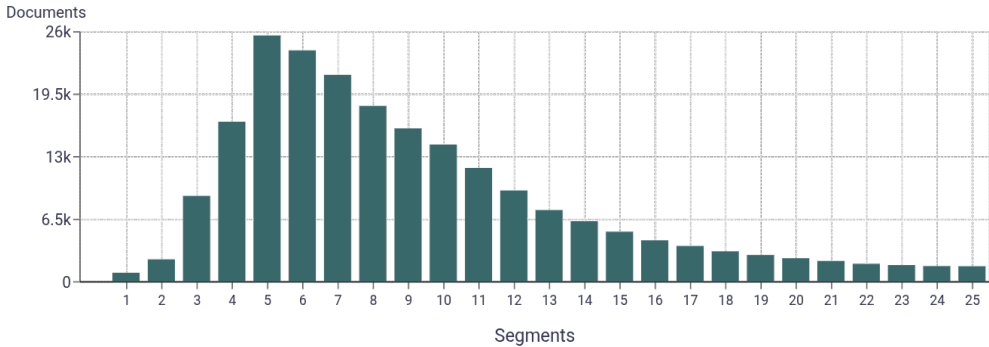
Top 10 domains

Domain	Docs	% of total
radiyoyacuvoo.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%

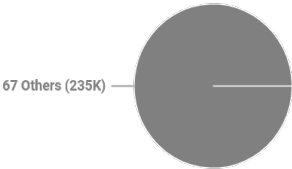
Documents size (in segments) ⓘ



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

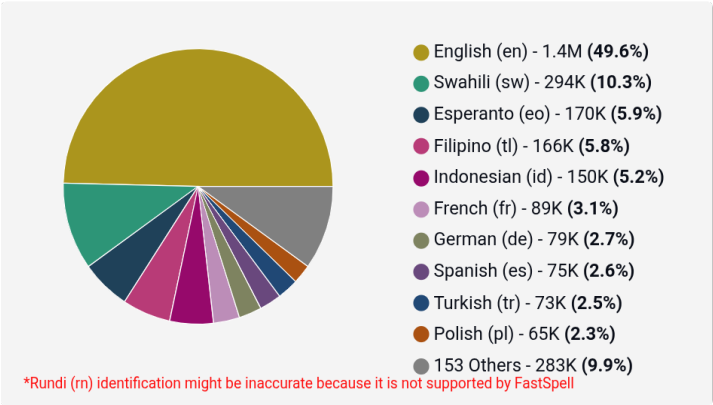
Document collections

CC = 89.94%
IA = 10.06%

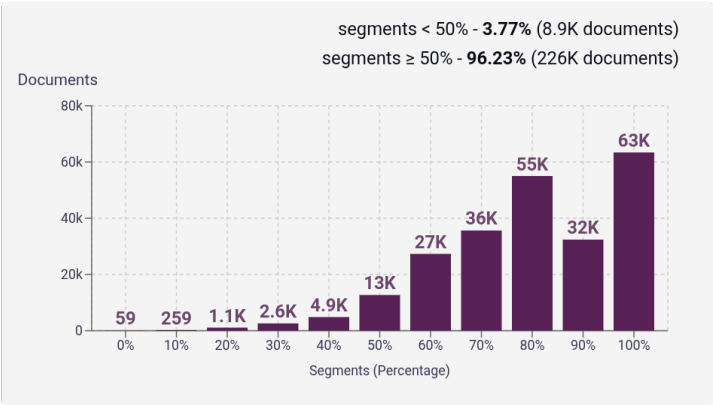


Language Distribution

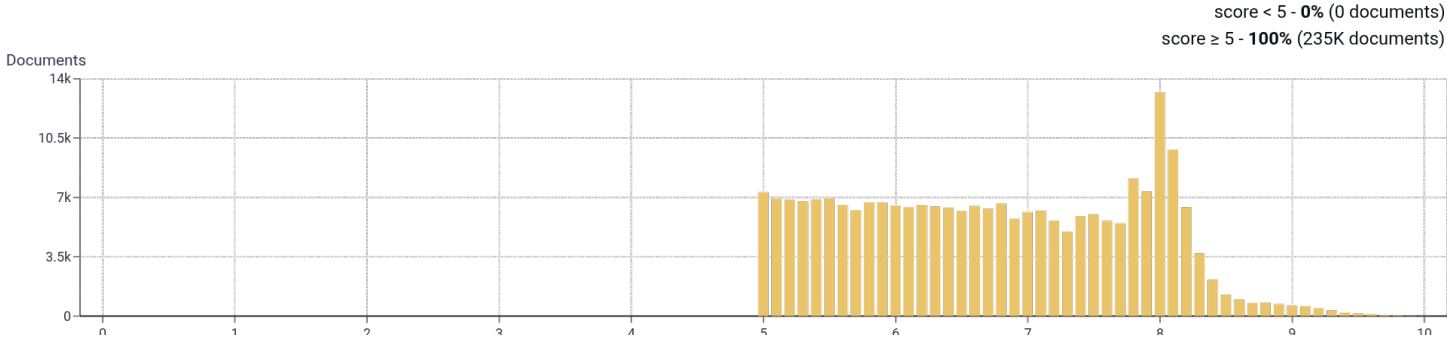
Number of segments in the Rundi (rn) corpus



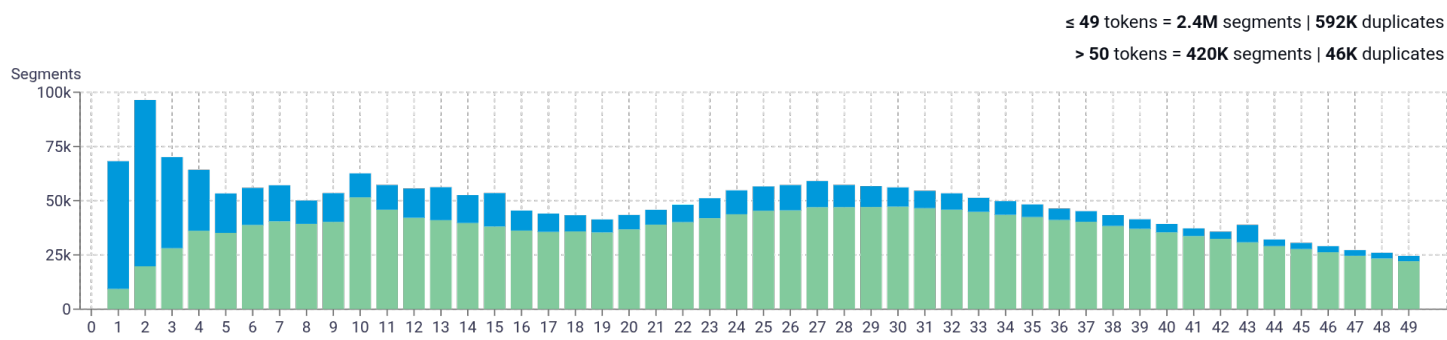
Percentage of segments in Rundi (rn) inside documents



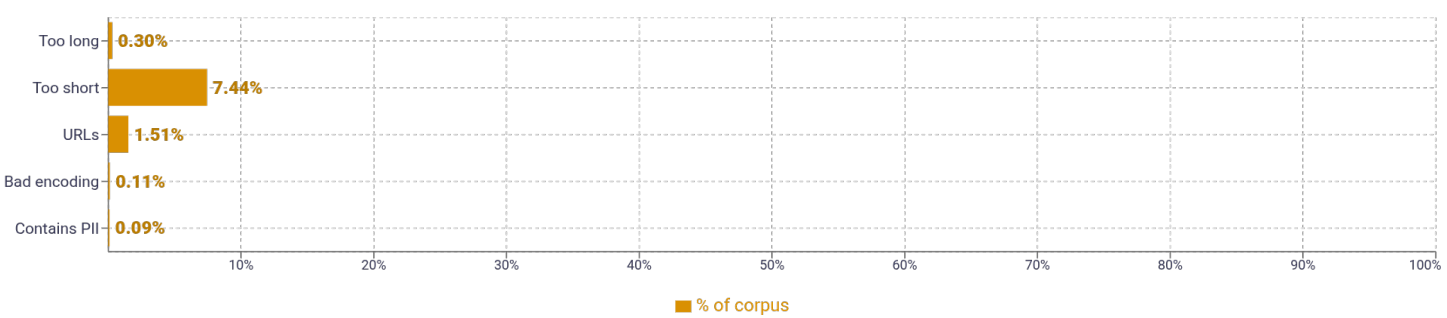
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				