

General overview

Corpus	Date	Language
hplt-v3-kin_Latn	9/18/2025	Kinyarwanda (rw)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
202,519	3,730,011	3,014,418 (80.82 %)	120M	690,017,631	669.92 MB

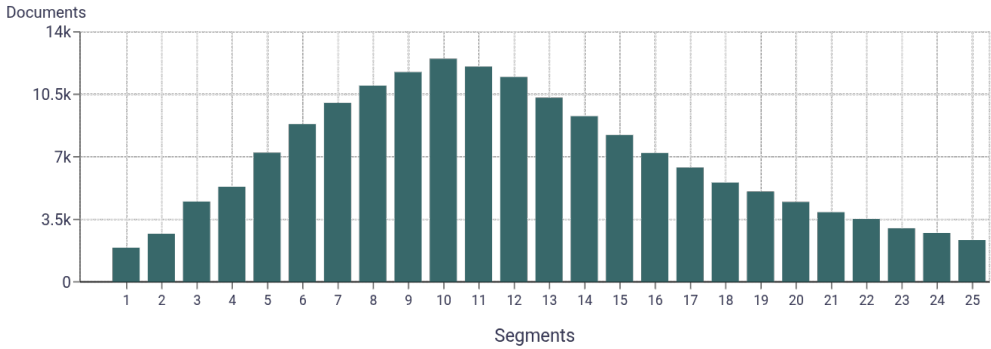
Top 10 domains

Domain	Docs	% of total
igihe.com	28K	13.92%
kigalitoday.com	15K	7.30%
inyarwanda.com	6.4K	3.17%
imvahonshya.co.rw	5.1K	2.51%
agakiza.org	4.7K	2.34%
umuryango.rw	4.6K	2.26%
jw.org	4K	1.96%
umuseke.rw	3.9K	1.95%
yegob.rw	3.2K	1.59%
yezu-akuzwe.org	3.1K	1.55%

Top 10 TLDs

Domain	Docs	% of total
com	127K	62.63%
rw	37K	18.49%
org	20K	10.05%
co.rw	7K	3.44%
gov.rw	2.8K	1.40%
net	2.8K	1.40%
fr	1.7K	0.83%
info	591	0.29%
ca	492	0.24%
xyz	386	0.19%

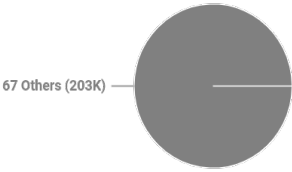
Documents size (in segments) ⓘ



≤ 25 segments **84.59%** (171K documents)  
> 25 segments **15.41%** (31K documents)

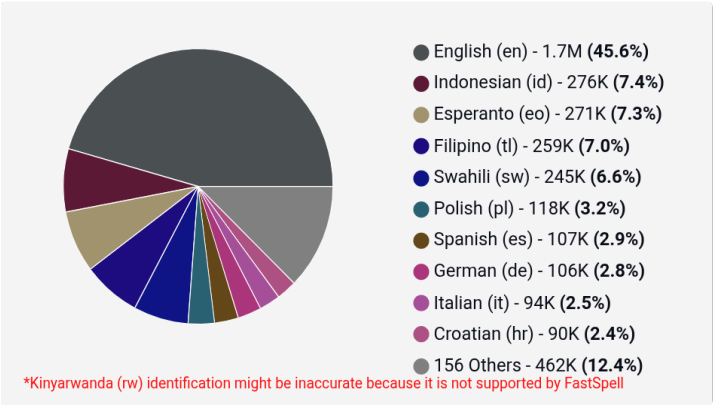
Document collections

CC = **88.45%**  
IA = **11.55%**

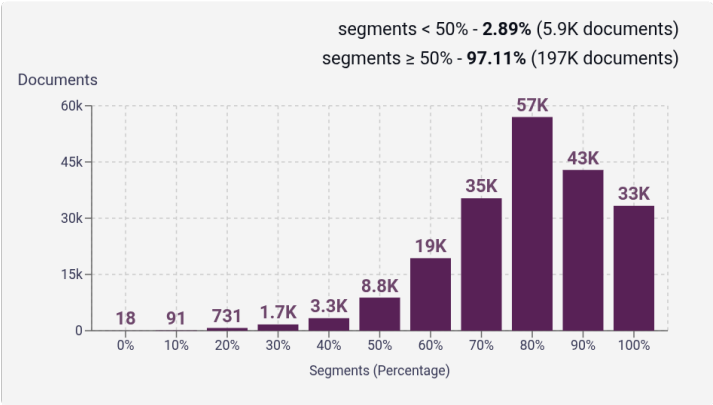


Language Distribution

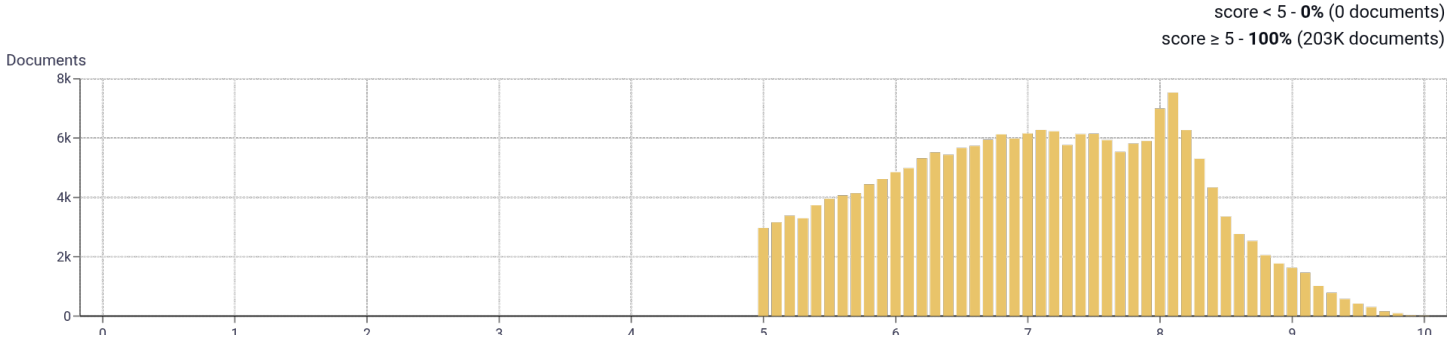
Number of segments in the Kinyarwanda (rw) corpus



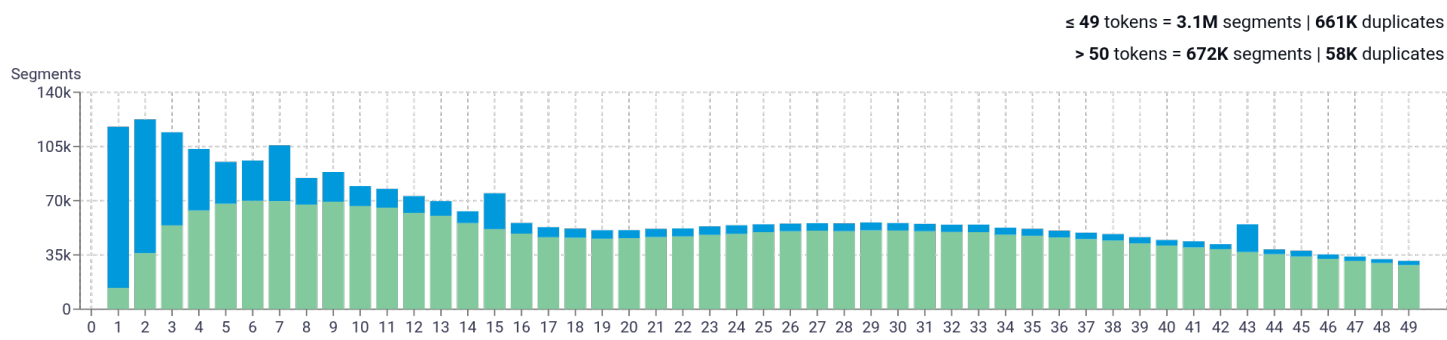
Percentage of segments in Kinyarwanda (rw) inside documents



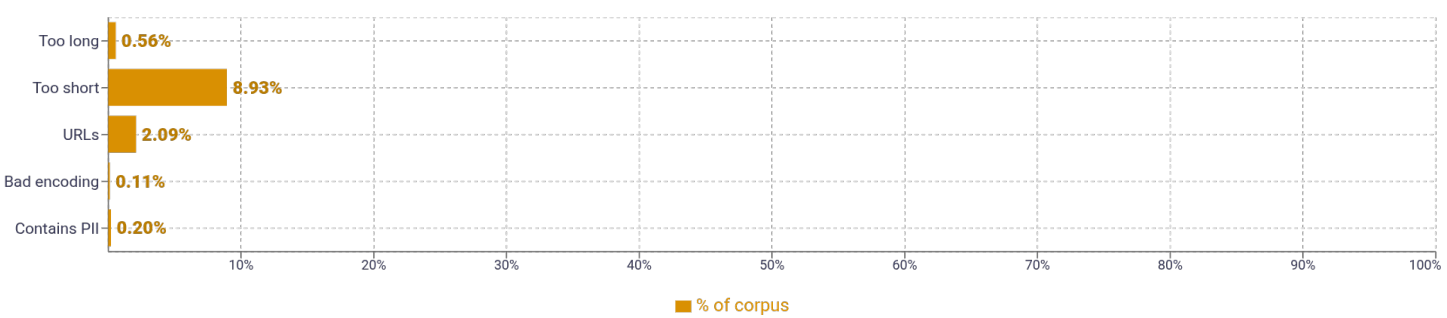
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>n   1,977,691</div> <div>y   855,626</div> <div>kandi   687,566</div> <div>w   460,481</div> <div>kuko   430,462</div>	
2	<div>u rwanda   117,822</div> <div>ndetse n   92,288</div> <div>nyuma y   69,343</div> <div>cyane cyane   41,979</div> <div>umuyobozi w   39,953</div>	
3	<div>kanda hano umusubize   27,789</div> <div>hirya no hino   17,850</div> <div>jenoside yakorewe abatutsi   17,831</div> <div>andika email yawe   14,854</div> <div>kutagaragara hano cyangwa   14,851</div>	
4	<div>kutagaragara hano cyangwa kigasibwa   14,851</div> <div>gishobora kutagaragara hano cyangwa   14,851</div> <div>isuzuma rikorwa na igihe   14,850</div> <div>igitekerezo cyawe kigaragara nyuma   14,850</div> <div>cyawe kigaragara nyuma y   14,850</div>	
5	<div>gishobora kutagaragara hano cyangwa kigasibwa   14,851</div> <div>igitekerezo cyawe kigaragara nyuma y   14,850</div> <div>igitekerezo cyanyu gishobora kutagaragara hano   14,849</div> <div>cyanyu gishobora kutagaragara hano cyangwa   14,849</div> <div>bidakurikijwe igitekerezo cyanyu gishobora kutagaragara   14,849</div>	

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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>, <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	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				