

General overview

Corpus	Date	Language
hplt-v3-knc_Latn	9/18/2025	Kanuri (kr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,387	30,472	21,138 (69.37 %)	1.7M	7,090,958	7.6 MB

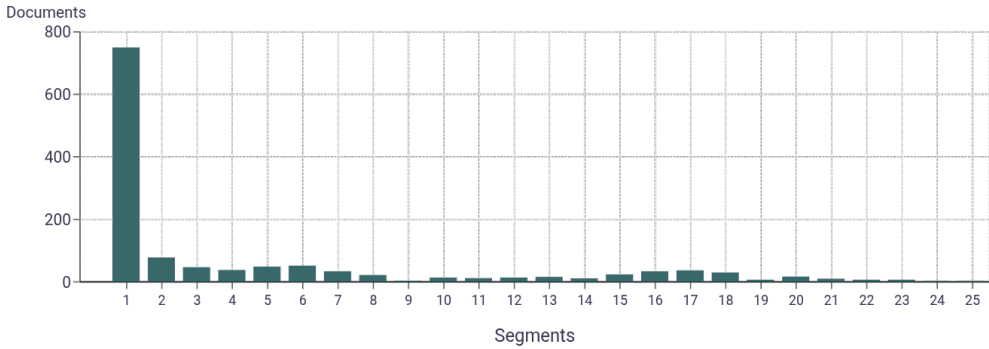
Top 10 domains

Domain	Docs	% of total
bible.is	755	54.43%
dandalkura.com	298	21.49%
wikimedia.org	121	8.72%
ebible.org	98	7.07%
fivecowries.online	45	3.24%
ngbible.com	11	0.79%
bibles.org	7	0.50%
bible.com	7	0.50%
boudouma.com	5	0.36%
shadowserver.org	4	0.29%

Top 10 TLDs

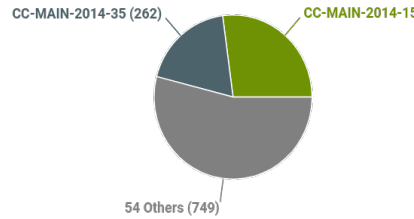
Domain	Docs	% of total
is	755	54.43%
com	333	24.01%
org	243	17.52%
online	45	3.24%
net	3	0.22%
ru	2	0.14%
club	2	0.14%
io	1	0.07%
info	1	0.07%
in	1	0.07%

Documents size (in segments) ⓘ



≤ 25 segments **95.17%** (1.3K documents)
> 25 segments **4.83%** (67 documents)

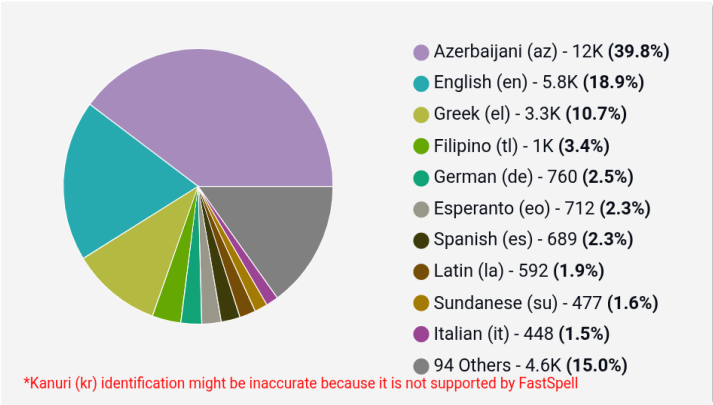
Document collections



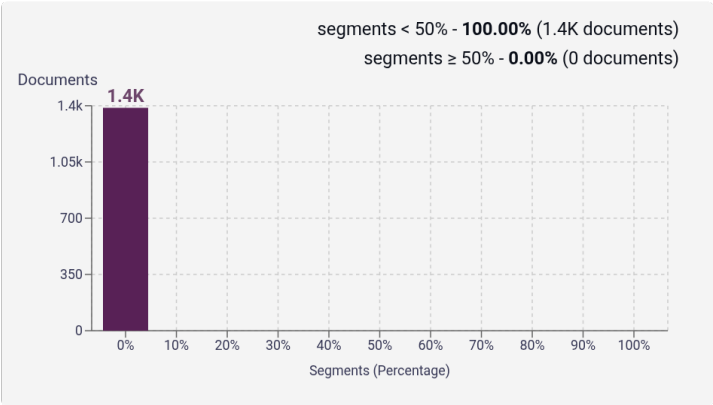
CC = 99.35%
IA = 0.65%

Language Distribution

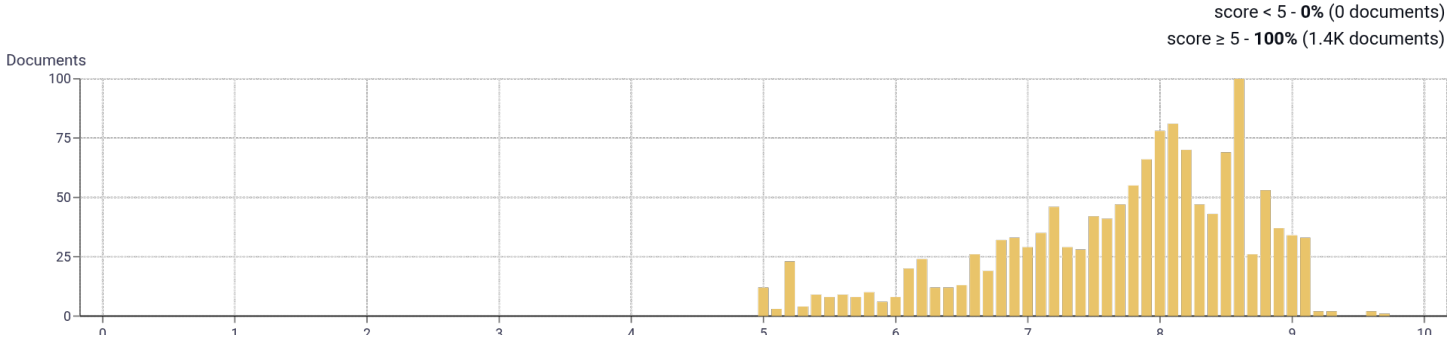
Number of segments in the Kanuri (kr) corpus



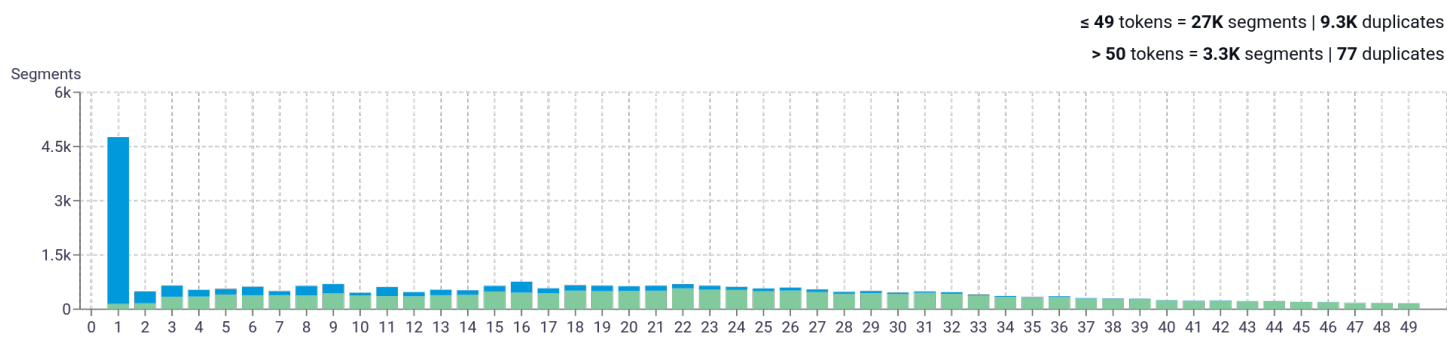
Percentage of segments in Kanuri (kr) inside documents



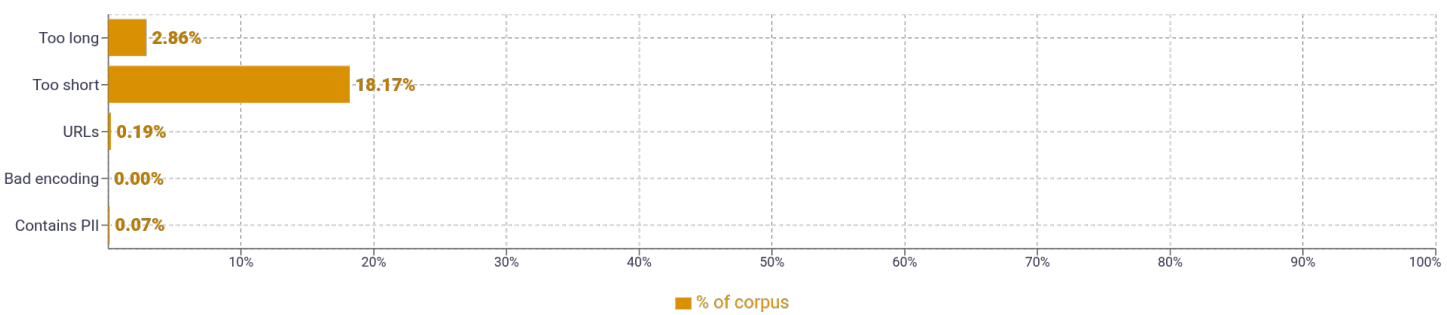
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>a 35,210</div> <div>ka 35,053</div> <div>ki 21,449</div> <div>na 21,129</div> <div>i 16,287</div>	
2	<div>ka ci 2,359</div> <div>ti ki 2,176</div> <div>ka tii 2,138</div> <div>tə ci 1,937</div> <div>je ki 1,838</div>	
3	<div>ane tuk na 734</div> <div>andza niye na 467</div> <div>bazlam i mbəlom 446</div> <div>a ndo hay 428</div> <div>a ba ka 384</div>	
4	<div>ha bazlam i mbəlom 206</div> <div>ndo məpe mədzal gər 201</div> <div>poy ta ki maji 198</div> <div>məpe mədzal gər hay 188</div> <div>mədzal gər hay ka 182</div>	
5	<div>məpe mədzal gər hay ka 182</div> <div>ndo məpe mədzal gər hay 179</div> <div>mədzal gər hay ka yesu 177</div> <div>məde ha bazlam i mbəlom 159</div> <div>ndo məde ha bazlam i 144</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				