

General overview

Corpus	Date	Language
hplt-v3-knc_Arab	9/18/2025	Kanuri (knc)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
912	26,735	25,047 (93.69 %)	1.7M	2,313,850	4.07 MB

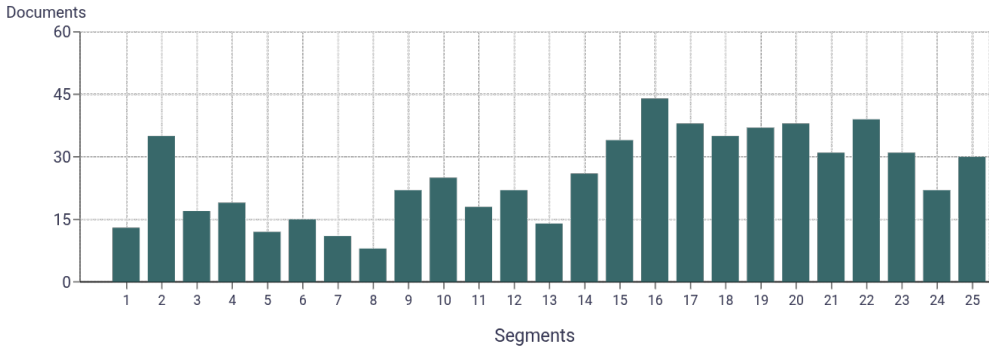
Top 10 domains

Domain	Docs	% of total
albayan.ae	220	24.12%
breakeveryyoke.com	122	13.38%
ebible.org	115	12.61%
bayan.id	39	4.28%
girlscoool.com	23	2.52%
biblearc.com	21	2.30%
folksocotra.org	17	1.86%
grorbnat.com	12	1.32%
blogfa.com	12	1.32%
asdika.org	12	1.32%

Top 10 TLDs

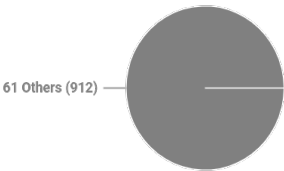
Domain	Docs	% of total
com	387	42.43%
ae	220	24.12%
org	167	18.31%
net	45	4.93%
id	41	4.50%
ws	9	0.99%
com.au	8	0.88%
ir	5	0.55%
de	5	0.55%
ru	4	0.44%

Documents size (in segments) ⓘ



≤ 25 segments **69.74%** (636 documents)
> 25 segments **30.26%** (276 documents)

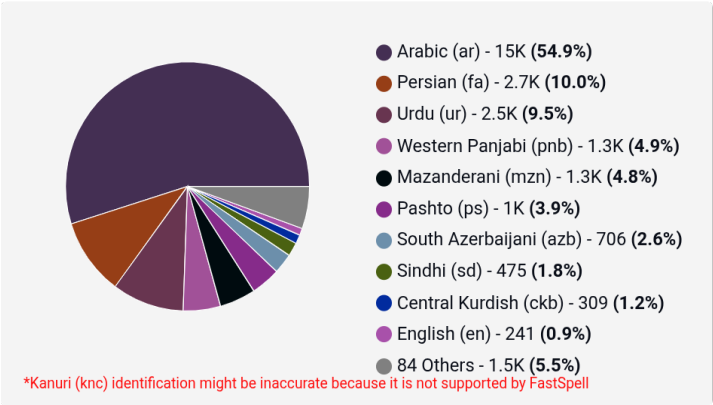
Document collections



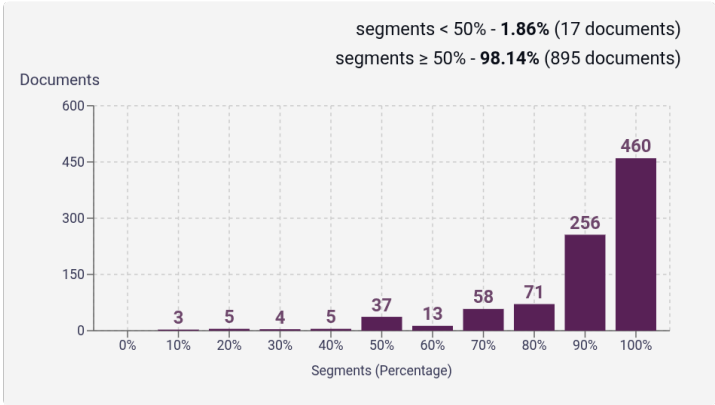
CC = 80.15%
IA = 19.85%

Language Distribution

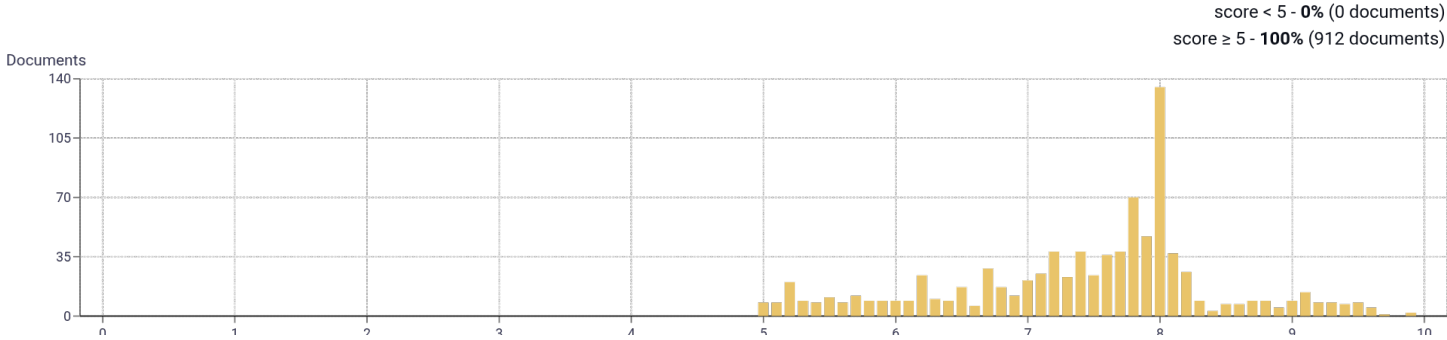
Number of segments in the Kanuri (knc) corpus



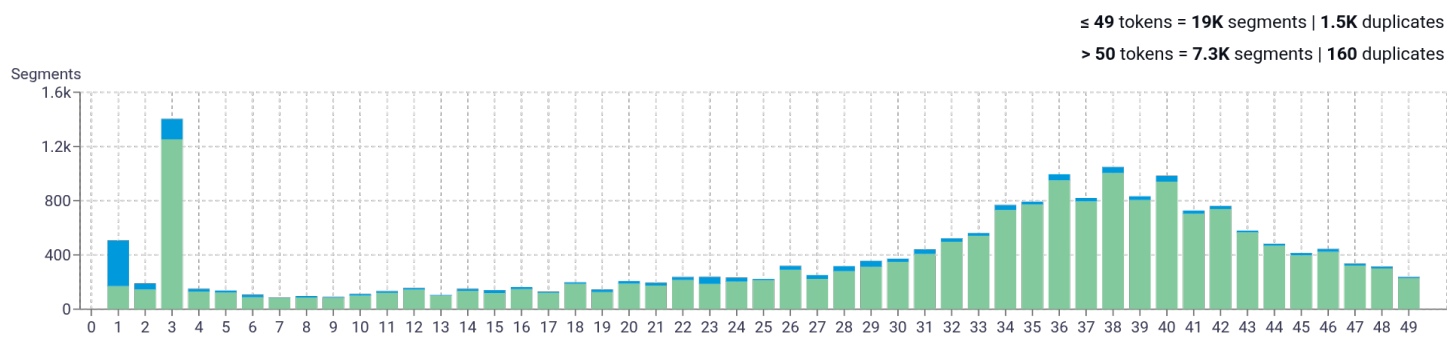
Percentage of segments in Kanuri (knc) inside documents



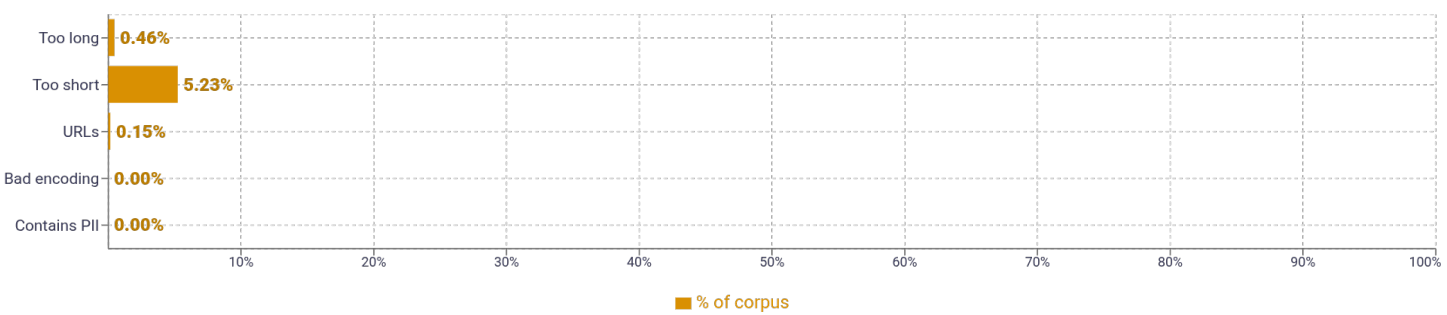
Distribution of documents by document score






Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>58,151 ي</div> <div>20,665 ه</div> <div>8,914 ي</div> <div>8,208 ك</div> <div>7,941 ال</div>	
2	<div>623 ب ن</div> <div>353 ن ي</div> <div>302 ك ن</div> <div>212 ي ي</div> <div>177 ال ي</div>	
3	<div>18 جامع المتون للحفظ</div> <div>10 - </div> <div>9 قراءة من الإنجيل</div> <div>9 الإنجيل بالدارجة المغربية</div> <div>8 منتديات غرور بنات</div>	
4	<div>9 قراءة من الإنجيل بالدارجة</div> <div>8 الموضوع الأصلي من هنا</div> <div>video courtesy of http 8</div> <div>some video courtesy of 8</div> <div>5 - </div>	
5	<div>9 قراءة من الإنجيل بالدارجة المغربية</div> <div>some video courtesy of http 8</div> <div>3 عبارات جميلة عن الحب والغرام</div> <div>3 المسجلين والمفعلين يمكنهم رؤية الوصلات</div> <div>الأعضاء المسجلين والمفعلين يمكنهم رؤية 3</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				