HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-knc_Latn	9/23/2025	Kanuri (knc)

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

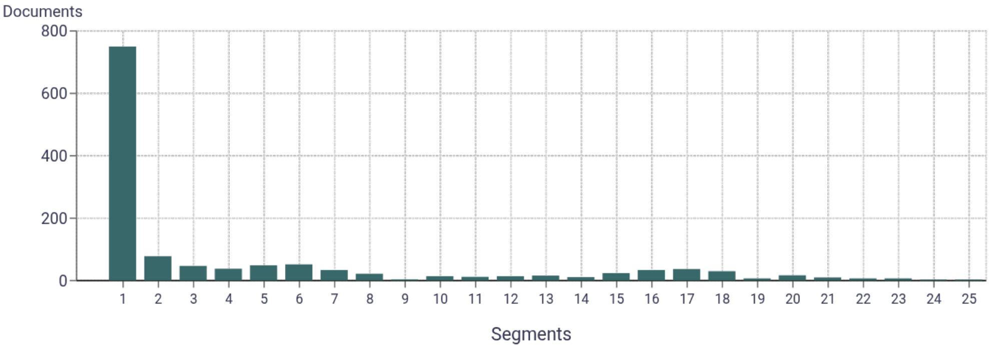
Domain	Docs	% of total
bible.is	755	54.43%
dandalkura.com	298	21.49%
wikimedia.org	121	8.72%
ebible.org	98	7.07%
fivecowries.online	45	3.24%
ngbible.com	11	0.79%
bibles.org	7	0.50%
bible.com	7	0.50%
boudouma.com	5	0.36%
shadowserver.org	4	0.29%

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%

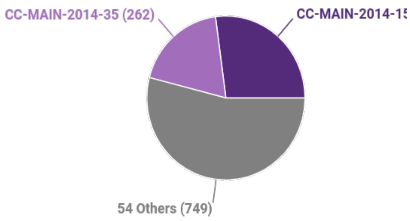
Documents size (in segments) ⓘ

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



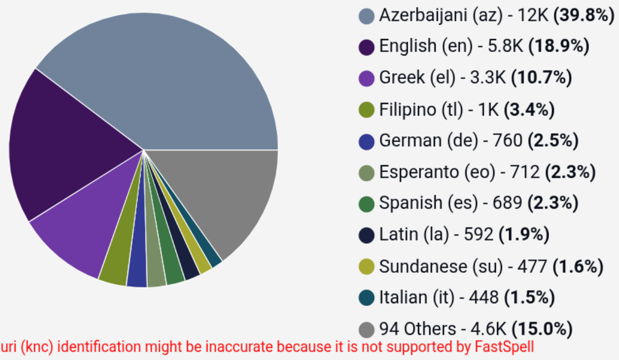
Document collections

CC = 99.35%
IA = 0.65%



Language Distribution

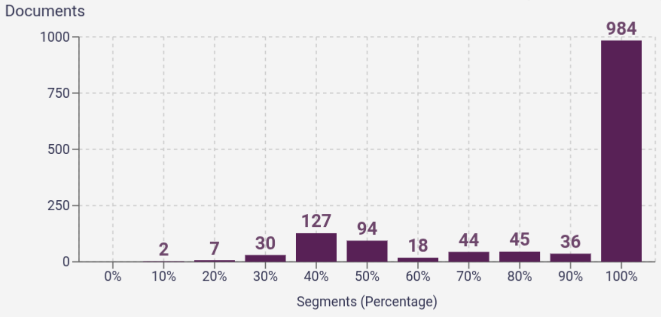
Number of segments in the Kanuri (knc) corpus



*Kanuri (knc) identification might be inaccurate because it is not supported by FastSpell

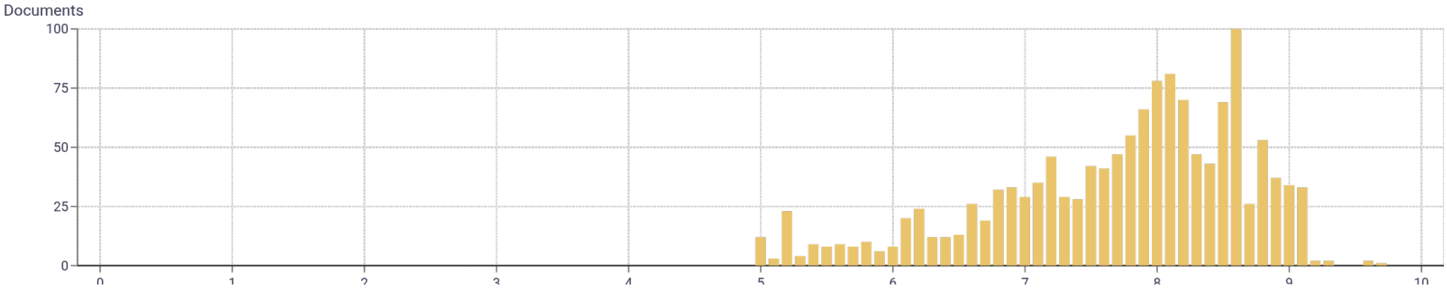
Percentage of segments in Kanuri (knc) inside documents

segments < 50% - **11.97%** (166 documents)
segments ≥ 50% - **88.03%** (1.2K documents)



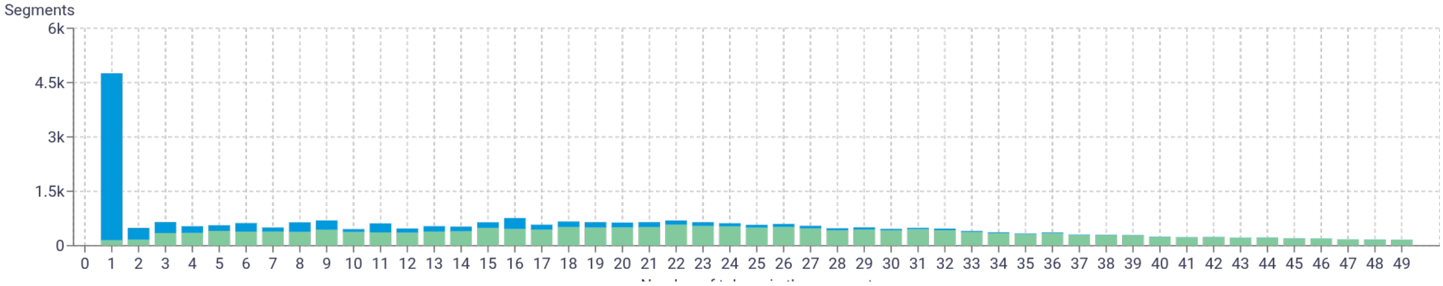
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.4K documents)

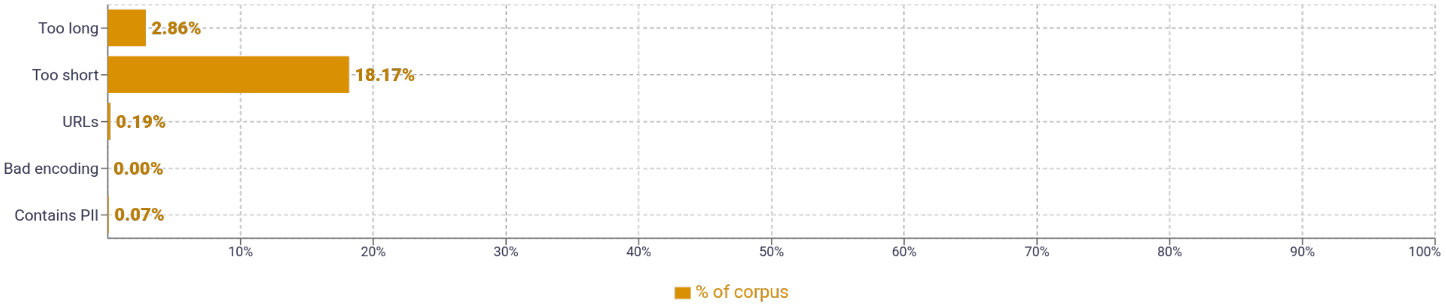


Segment length distribution by token

≤ **49** tokens = **27K** segments | **9.3K** duplicates
> **50** tokens = **3.3K** segments | **77** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>SIZE</th><th>N-GRAMS</th></tr>
<tr><td>1</td><td>ka | 35,053
kɨ | 21,449
na | 21,129
tə | 15,646
tɨ | 14,222</td></tr>
<tr><td>2</td><td>tɨ kɨ | 2,176
ka tii | 2,138
je kɨ | 1,838
aye na | 1,754
ə nə | 1,732</td></tr>
<tr><td>3</td><td>ane tuk na | 734
andza niye na | 467
bazlam i mbəlom | 446
əŋki ci ka | 344
taa wu patə | 315</td></tr>
<tr><td>4</td><td>ha bazlam i mbəlom | 206
ndo məpe mədzal gər | 201
poy ta kɨ majɨ | 198
məpe mədzal gər hay | 188
mədzal gər hay ka | 182</td></tr>
<tr><td>5</td><td>məpe mədzal gər hay ka | 182
ndo məpe mədzal gər hay | 179
mədzal gər hay ka yesu | 177
məɗe ha bazlam i mbəlom | 159
o fal kanuri milion mew | 142</td></tr>
</table>

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av