HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-dik_Latn	9/18/2025	Dinka (dik)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,223	32,639	29,330 (89.86 %)	1.7M	6,635,520	7.33 MB

Top 10 domains

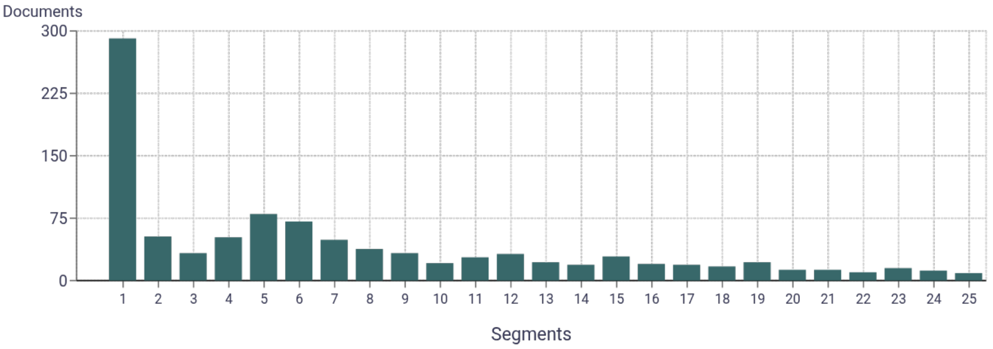
Domain	Docs	% of total
bible.is	296	24.20%
wikipedia.org	267	21.83%
ebible.org	172	14.06%
pngscriptures.org	66	5.40%
png.bible	45	3.68%
sbs.com.au	34	2.78%
breakeveryyoke.com	33	2.70%
stepbible.org	27	2.21%
bringingupgreat...	17	1.39%
www.vic.gov.au	16	1.31%

Top 10 TLDs

Domain	Docs	% of total
org	594	48.57%
is	296	24.20%
vic.gov.au	73	5.97%
com	65	5.31%
bible	45	3.68%
gov.au	42	3.43%
com.au	41	3.35%
org.au	32	2.62%
net.au	11	0.90%
vn	6	0.49%

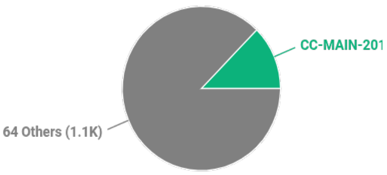
Documents size (in segments) ⓘ

≤ 25 segments **81.85%** (1K documents)
> 25 segments **18.15%** (222 documents)



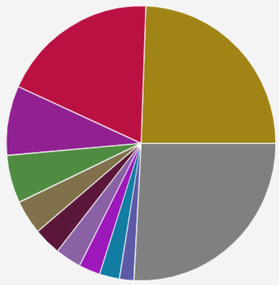
Document collections

CC = 94.60%
IA = 5.40%



Language Distribution

Number of segments in the Dinka (dik) corpus

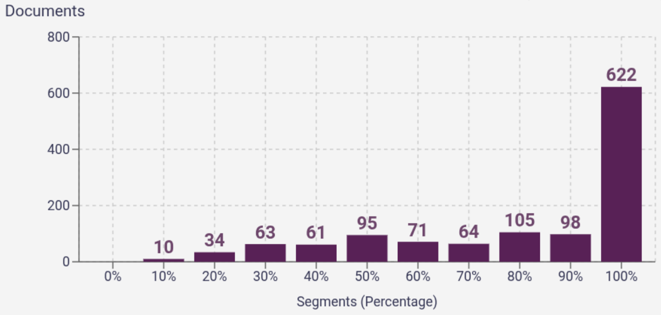


- ● Albanian (sq) - 8K **(24.4%)**
- ● English (en) - 6.1K **(18.8%)**
- ● Finnish (fi) - 2.7K **(8.2%)**
- ● Slovenian (sl) - 1.8K **(5.7%)**
- ● Sundanese (su) - 1.3K **(4.0%)**
- ● French (fr) - 1.1K **(3.3%)**
- ● Japanese (ja) - 1K **(3.1%)**
- ● Malay (ms) - 839 **(2.6%)**
- ● Dutch (nl) - 782 **(2.4%)**
- ● Romanian (ro) - 563 **(1.7%)**
- ● 123 Others - 8.4K **(25.8%)**

*Dinka (dik) identification might be inaccurate because it is not supported by FastSpell

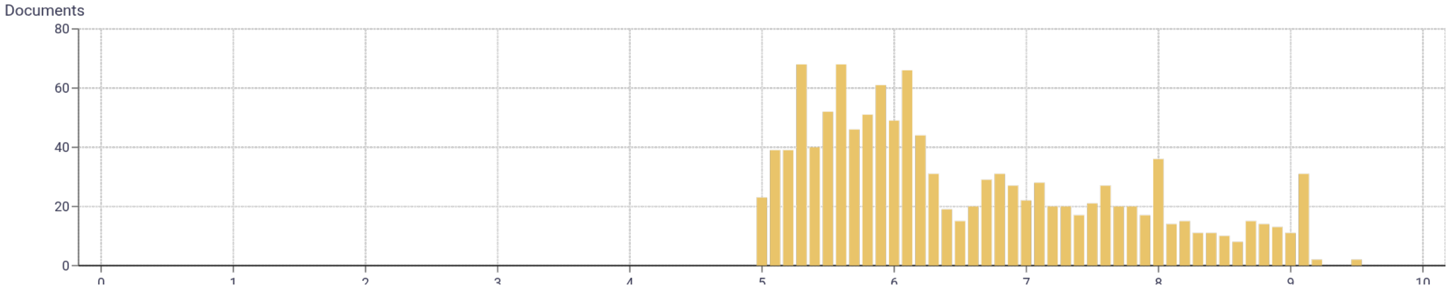
Percentage of segments in Dinka (dik) inside documents

segments < 50% - **13.74%** (168 documents)
segments ≥ 50% - **86.26%** (1.1K documents)



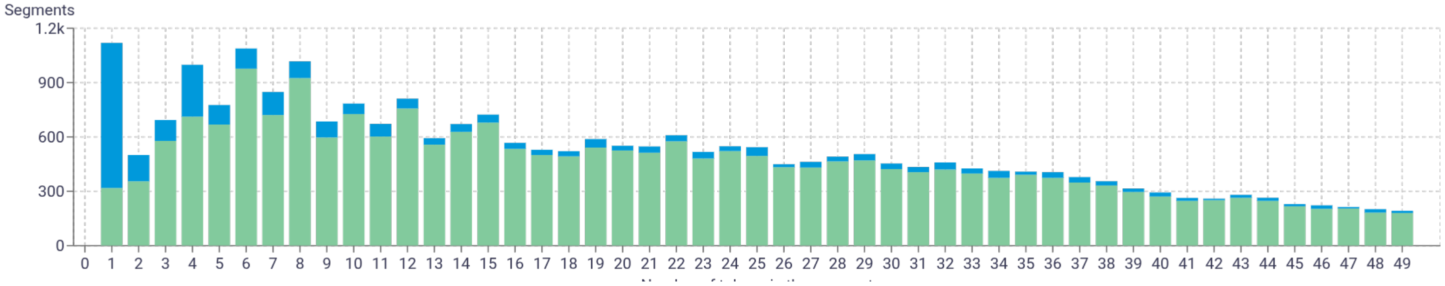
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.2K documents)

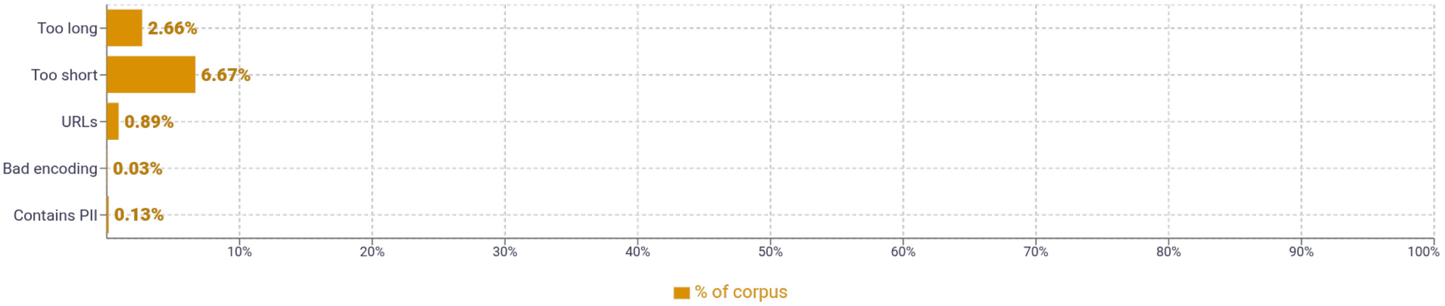


Segment length distribution by token

≤ **49** tokens = **26K** segments | **3.1K** duplicates
> **50** tokens = **6.6K** segments | **249** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>SIZE</th><th>N-GRAMS</th></tr>
<tr><td>1</td><td>ë | 18,703
kek | 14,345
raan | 12,846
nhialic | 10,927
yen | 10,541</td></tr>
<tr><td>2</td><td>to kek | 3,541
kek to | 3,541
southwestern dinka | 998
wek aa | 818
käk nhialic | 676</td></tr>
<tr><td>3</td><td>lɔc ku dɔc | 980
raan cï lɔc | 965
t puɔth yam | 337
kɔcken ye buɔɔth | 329
cie kɔc itharel | 280</td></tr>
<tr><td>4</td><td>athör thɛɛr wël nhialic | 252
akut kɔc cï gam | 183
tim cï rïïu kɔ | 157
ë thäät ë thäät | 143
thɛɛr wël nhialic yic | 140</td></tr>
<tr><td>5</td><td>raan cï lɔc ku dɔc | 959
athör thɛɛr wël nhialic yic | 140
t athör thɛɛr wël nhialic | 124
thäät ë thäät ë thäät | 106
ë thäät ë thäät ë | 105</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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
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				