

General overview

Corpus	Date	Language
hplt-v3-dik_Latn	9/18/2025	Southwestern Dinka

Volumes

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

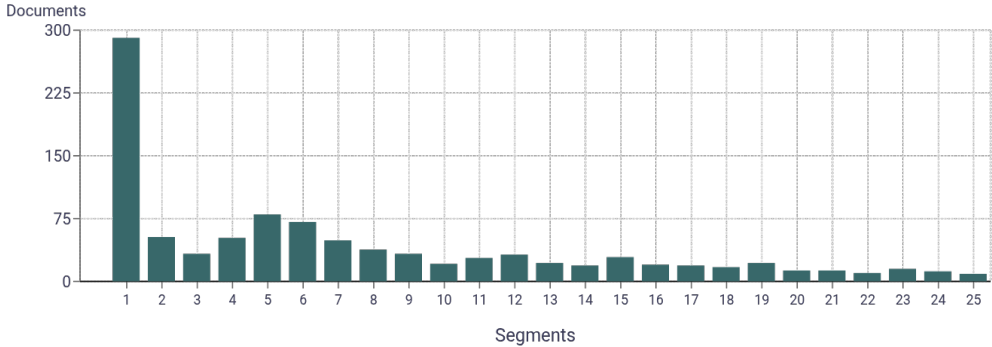
Top 10 domains

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

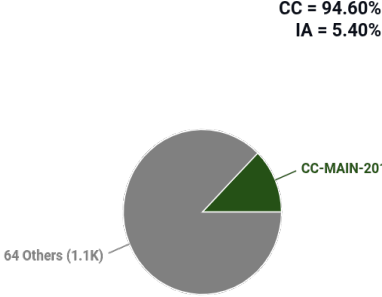
Domain	Docs	% of total
org	594	48.57%
is	296	24.20%
vic.gov.au	73	5.97%
com	65	5.31%
bible	45	3.68%
gov.au	42	3.43%
com.au	41	3.35%
org.au	32	2.62%
net.au	11	0.90%
vn	6	0.49%

Documents size (in segments) ⓘ



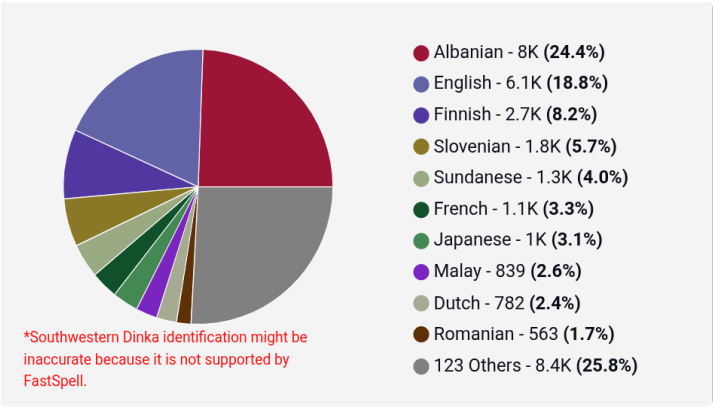
≤ 25 segments **81.85%** (1K documents)
> 25 segments **18.15%** (222 documents)

Document collections

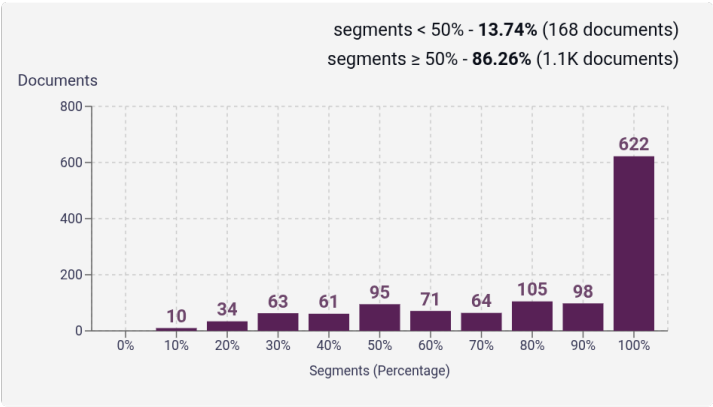


Language Distribution

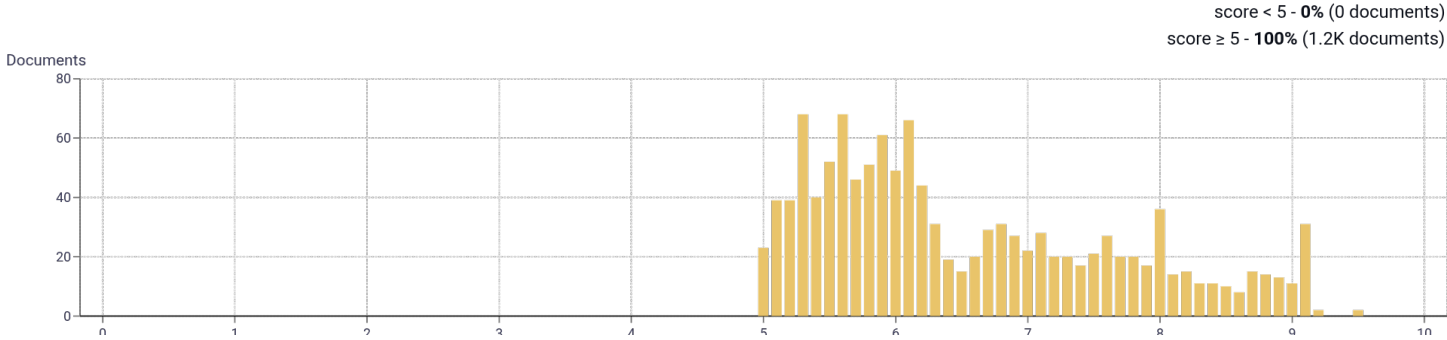
Number of segments in the Southwestern Dinka corpus



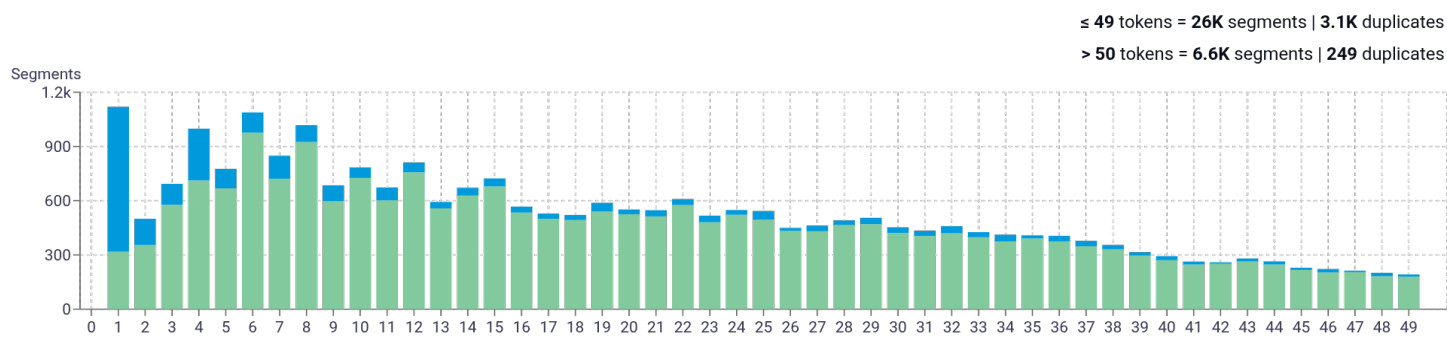
Percentage of segments in Southwestern Dinka inside documents



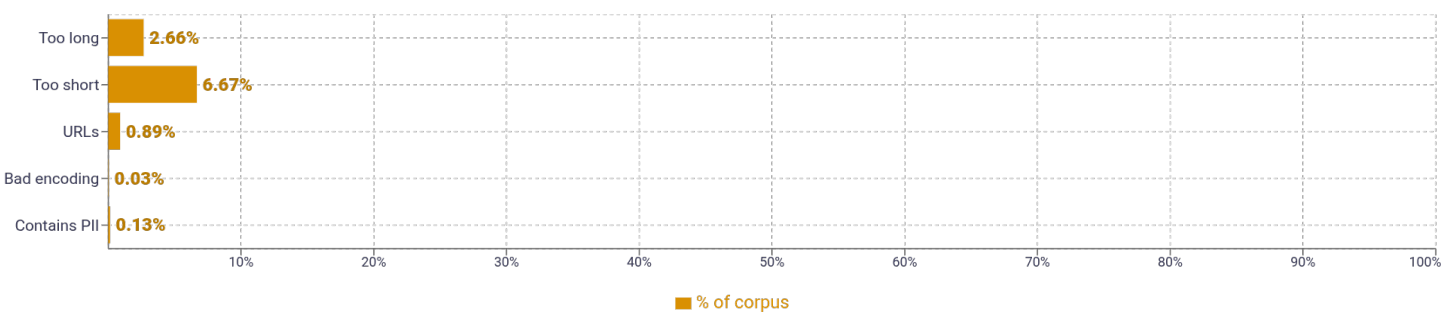
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>ë 18,703</div> <div>kek 14,345</div> <div>raan 12,846</div> <div>nhialic 10,927</div> <div>yen 10,541</div>	
2	<div>to kek 3,541</div> <div>kek to 3,541</div> <div>southwestern dinka 998</div> <div>wek aa 818</div> <div>käk nhialic 676</div>	
3	<div>bc ku dɔc 980</div> <div>raan cĩ bc 965</div> <div>t puɔth yam 337</div> <div>kɔcken ye buɔth 329</div> <div>cie kɔc ithare! 280</div>	
4	<div>athør theer wēl nhialic 252</div> <div>akut kɔc cĩ gam 183</div> <div>tĩm cĩ rĩlu kɔ 157</div> <div>ë thăät ë thăät 143</div> <div>theer wēl nhialic yic 140</div>	
5	<div>raan cĩ bc ku dɔc 959</div> <div>athør theer wēl nhialic yic 140</div> <div>t athør theer wēl nhialic 124</div> <div>thăät ë thăät ë thăät 106</div> <div>ë thăät ë thăät ë 105</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				