

General overview

Corpus	Date	Language
hplt-v3-fij_Latn	9/18/2025	Fijian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
12,071	283,499	224,456 (79.17 %)	13M	59,123,262	56.73 MB

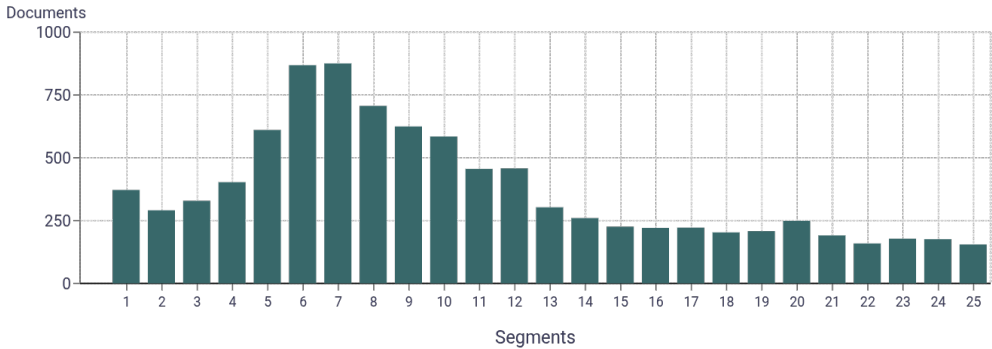
Top 10 domains

Domain	Docs	% of total
jw.org	3.3K	27.44%
vitifm.com.fj	2.3K	18.76%
fijitimes.com	1.1K	8.74%
wordproject.org	1K	8.62%
churchofjesusch...	544	4.51%
wikipedia.org	307	2.54%
bible.is	262	2.17%
fijilive.com	260	2.15%
bibles.org	158	1.31%
lds.org	154	1.28%

Top 10 TLDs

Domain	Docs	% of total
org	5.8K	47.65%
com	3K	24.44%
com.fj	2.5K	20.84%
is	262	2.17%
govt.nz	97	0.80%
com.au	60	0.50%
gov.fj	54	0.45%
net	39	0.32%
co	30	0.25%
edu.pl	20	0.17%

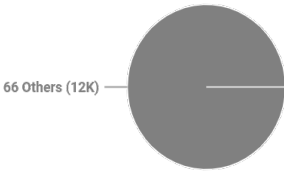
Documents size (in segments) ⓘ



≤ 25 segments 77.32% (9.3K documents)
> 25 segments 22.68% (2.7K documents)

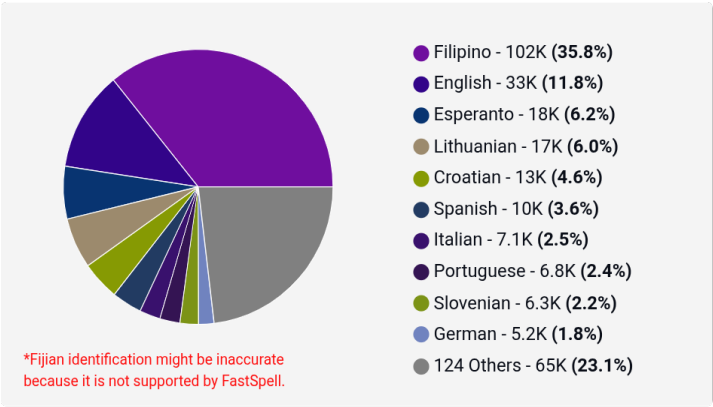
Document collections

CC = 91.24%
IA = 8.76%

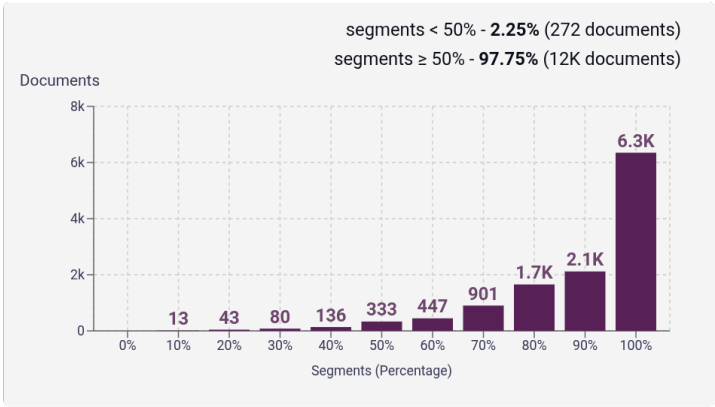


Language Distribution

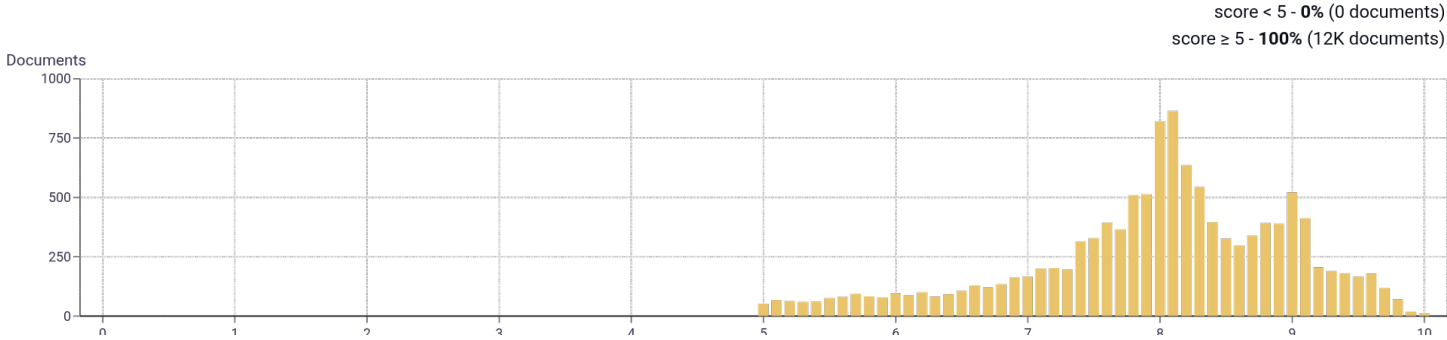
Number of segments in the Fijian corpus



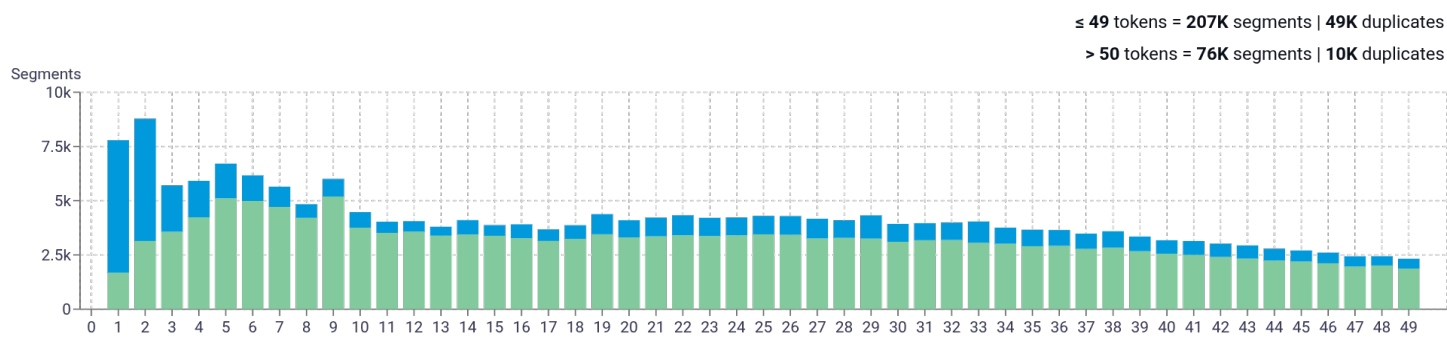
Percentage of segments in Fijian inside documents



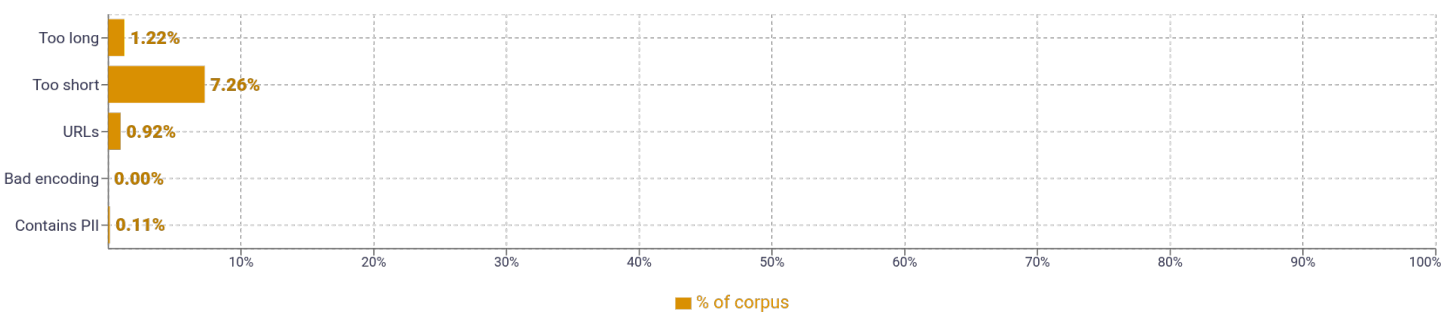
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	kina 113,050 dua 103,196 ga 96,611 i 96,135 era 90,883	
2	tale ga 19,210 i jiova 11,220 sara ga 8,354 mada ga 4,893 dua tale 4,822	
3	kina e dua 3,163 dua na gauna 2,833 dua na tamata 2,538 matanitu ni kalou 2,150 vosa ni kalou 2,138	
4	jiova ni lewe vuqa 676 tagane kei na yalewa 610 jiova na nomuni kalou 555 i cabocabo ni soro 487 soro ni valavala ca 483	
5	dia ta lala dia kata 293 written by mika qalobula on 291 wili ivolatabu ni macawa qo 281 vakaoqo na kalou o jiova 269 tucake tale mai na mate 260	

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				