

General overview

Corpus	Date	Language
hplt-v3-tpi_Latn	9/18/2025	Tok Pisin

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
12,425	266,498	229,043 (85.95 %)	13M	58,138,560	55.83 MB

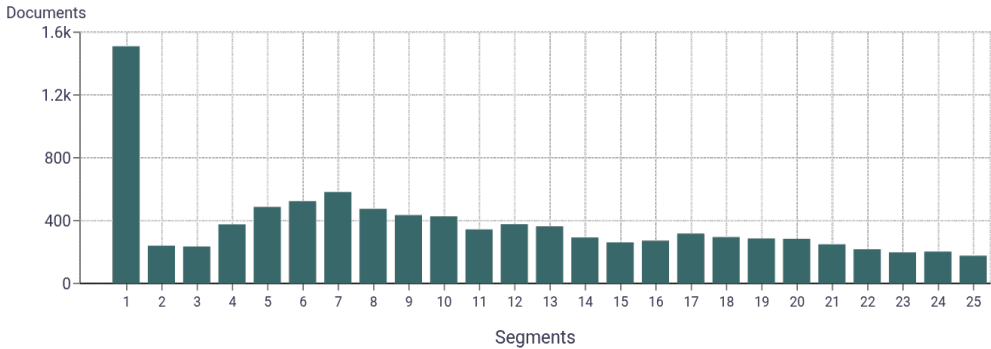
Top 10 domains

Domain	Docs	% of total
jw.org	3.3K	26.40%
vbtc.vu	1.8K	14.49%
bible.is	1.4K	11.62%
ebible.org	998	8.03%
abc.net.au	580	4.67%
aboriginalbible...	465	3.74%
pngbiblesourc...	412	3.32%
gov.vu	244	1.96%
globalrecording...	232	1.87%
emtv.com.pg	205	1.65%

Top 10 TLDs

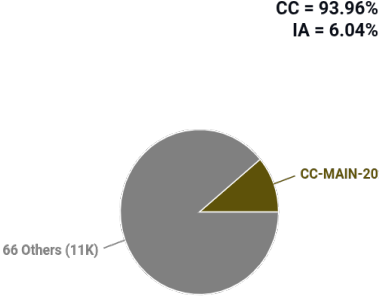
Domain	Docs	% of total
org	5.6K	44.71%
vu	2.4K	19.03%
is	1.4K	11.65%
com	891	7.17%
net.au	664	5.34%
org.au	476	3.83%
net	288	2.32%
com.pg	228	1.84%
us	80	0.64%
info	73	0.59%

Documents size (in segments) ⓘ



≤ 25 segments **75.85%** (9.4K documents)
> 25 segments **24.15%** (3K documents)

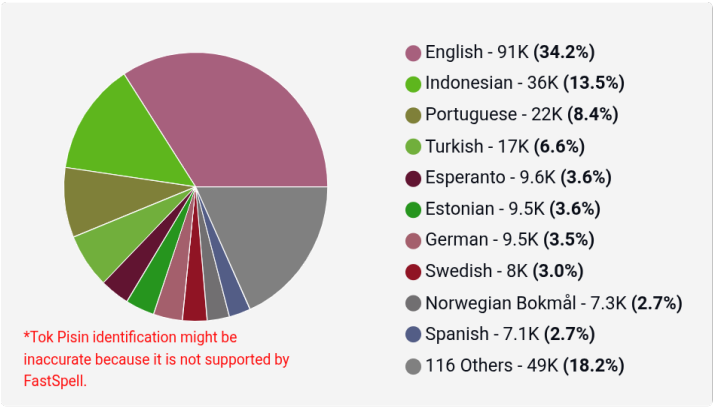
Document collections



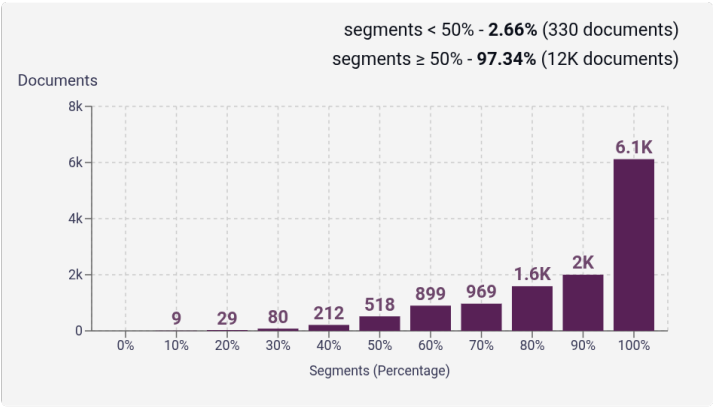
CC = 93.96%
IA = 6.04%

Language Distribution

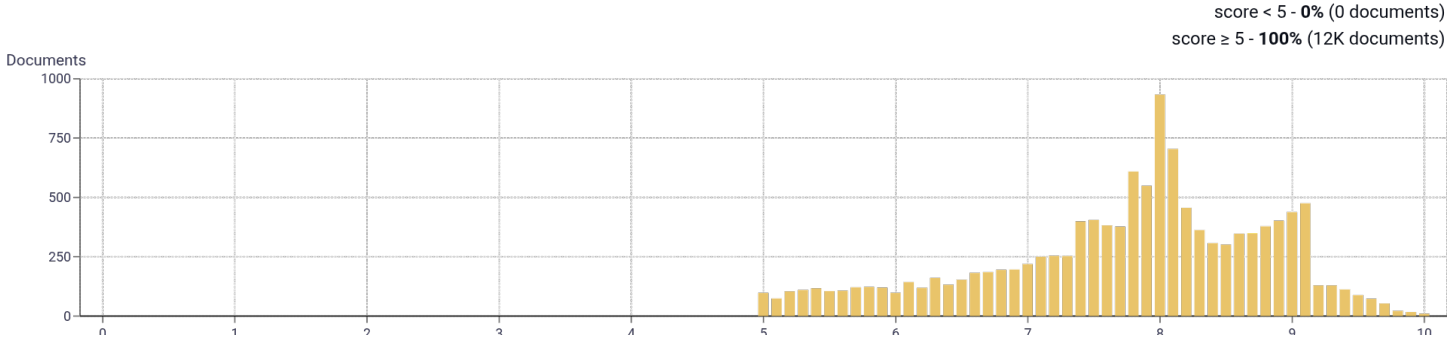
Number of segments in the Tok Pisin corpus



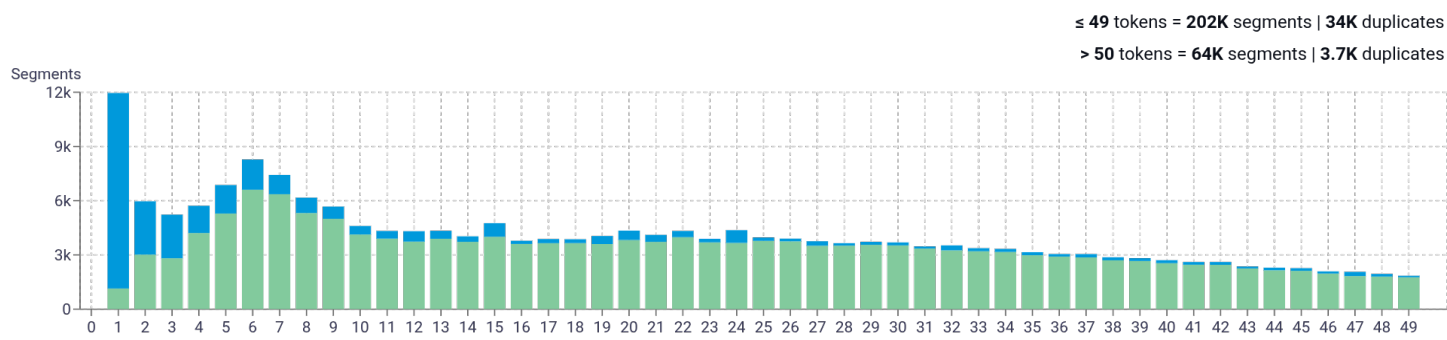
Percentage of segments in Tok Pisin inside documents



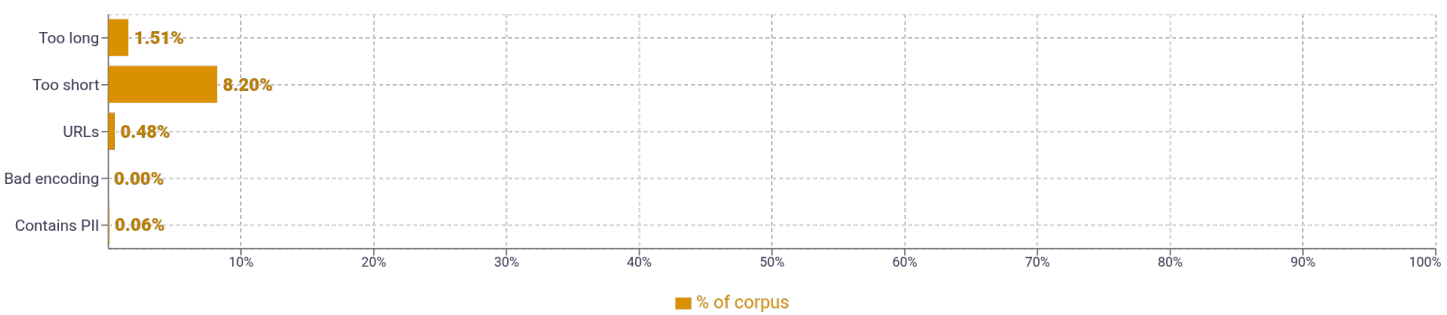
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>mi 136,306</div> <div>yu 130,143</div> <div>hem 128,571</div> <div>we 99,764</div> <div>god 91,288</div>	
2	<div>we oli 19,773</div> <div>talem se 15,919</div> <div>langa det 7,176</div> <div>bai mi 6,832</div> <div>yu mas 6,493</div>	
3	<div>yupela i mas 4,252</div> <div>hem i talem 2,528</div> <div>yu no ken 1,938</div> <div>hemi talem se 1,815</div> <div>yufala i mas 1,750</div>	
4	<div>yupela i no ken 1,773</div> <div>feitful wok bilong jisas 1,438</div> <div>hem i talem se 1,391</div> <div>bai i no inap 985</div> <div>baebol i talem se 857</div>	
5	<div>we i no save finis 531</div> <div>up and down arrows for 500</div> <div>space to play or pause 500</div> <div>trastim god long kisim bek 402</div> <div>ofa bilong paia i kukim 302</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				