HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
tpi_Latn.jsonl.tsv	11/27/2024	Tok Pisin (tpi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
13,978	282,367	147,572 (52.26 %)	14M	61.65 MB	64,257,063

Top 10 domains

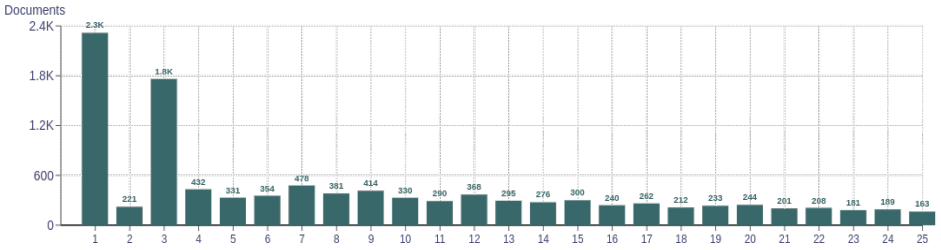
Domain	Docs	% of total
bible.is	3.6K	25.44
jw.org	2.5K	17.63
vbtc.vu	1.4K	10.27
pngscriptures.org	1.1K	8.00
wikipedia.org	1K	7.33
aboriginalbibles.org.au	490	3.51
pngbibleresources.org	452	3.23
ebible.org	246	1.76
png.bible	216	1.55
breakeveryyoke.com	199	1.42

Top 10 TLDs

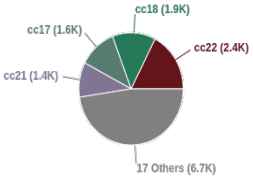
Domain	Docs	% of total
org	5.8K	41.33
is	3.6K	25.45
vu	1.8K	13.08
com	1.1K	7.52
org.au	512	3.66
bible	216	1.55
net.au	210	1.50
net	167	1.19
com.pg	148	1.06
us	120	0.86

Documents size (in segments)

<= 25 segments **76.41%** (11K documents)
> 25 segments **23.59%** (3.3K documents)

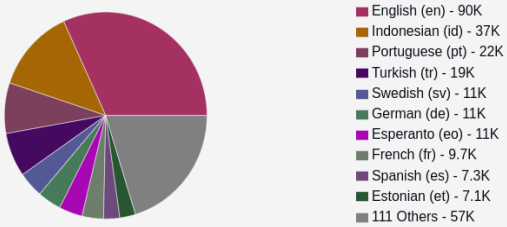


Documents by collection



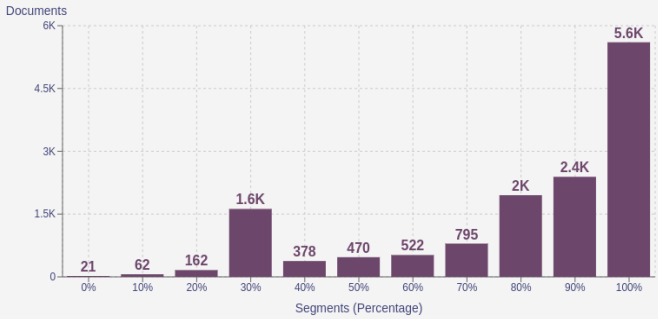
Language Distribution

Number of segments



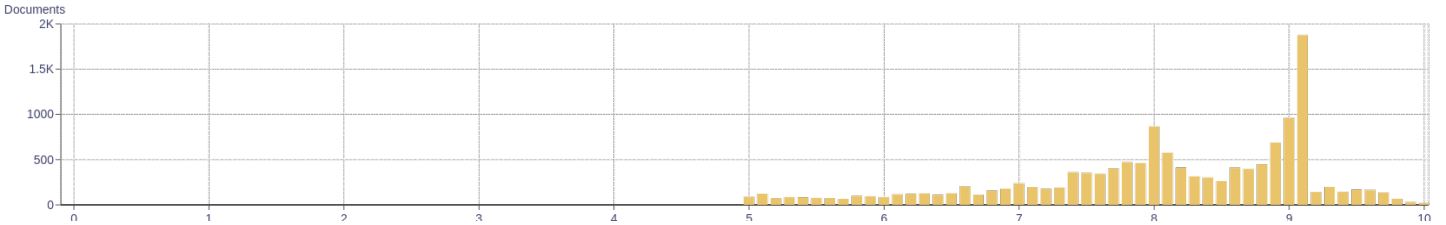
*Tok Pisin (tpi) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Tok Pisin (tpi) inside documents



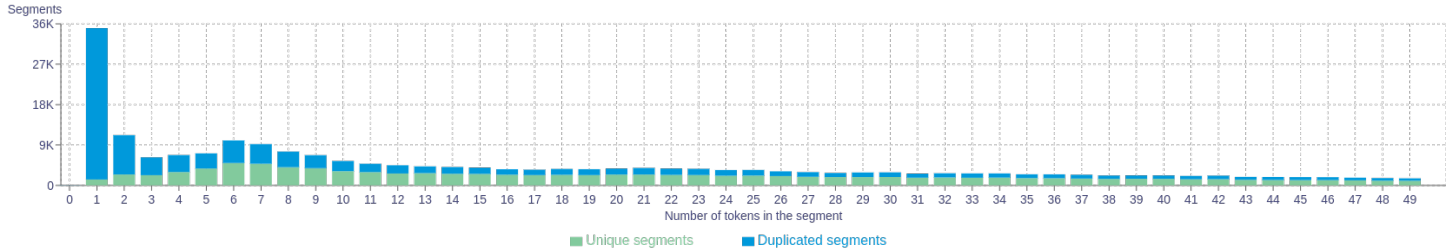
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (14K documents)

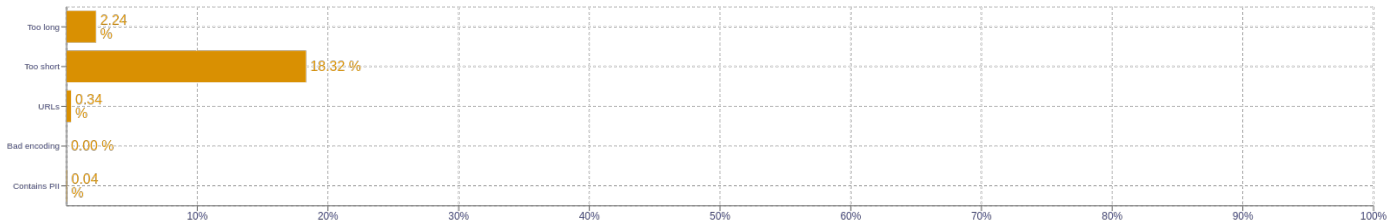


Segment length distribution by token

<= 49 tokens = **107K** segments | **113K** duplicates
> 50 tokens = **63K** segments | **22K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	mi | 183796 · yu | 159610 · hem | 124839 · god | 119262 · we | 108100
2	we oli | 19692 · talem se | 16176 · bai mi | 12004 · god god | 11132 · yu mas | 8025
3	god god god | 11112 · yupela i mas | 7785 · yufala i mas | 2891 · hem i talem | 2805 · yu no ken | 2654
4	god god god god | 11092 · yupela i no ken | 3039 · bai i no inap | 1618 · hem i talem se | 1518 · feitful wok bilong jisas | 1334
5	god god god god god | 11072 · we i no save finis | 765 · flap flap flap flap flap | 738 · ofa bilong paia i kukim | 645 · smok i gat gutpela smel | 501

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