HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bjn_Arab	9/17/2025	Banjar

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,306	30,400	24,867 (81.80 %)	1.1M	4,582,094	7.88 MB

Top 10 domains

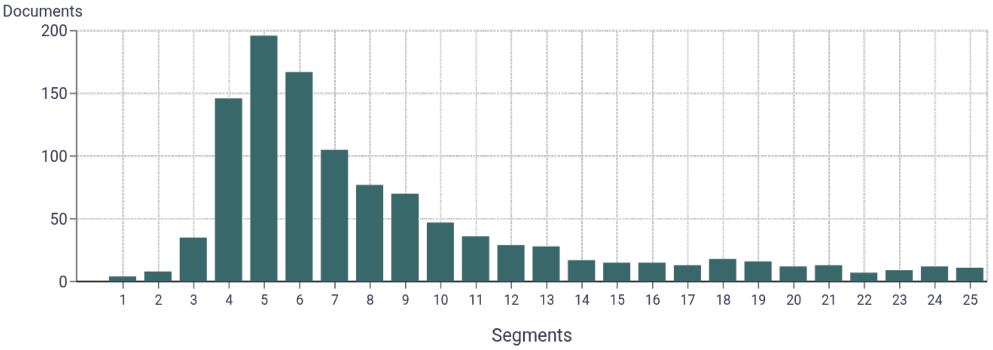
Domain	Docs	% of total
utusanmelayu.co...	662	50.69%
blogspot.com	153	11.72%
wordpress.com	147	11.26%
ahmadalikarim.com	54	4.13%
utusantv.com	33	2.53%
wikimedia.org	19	1.45%
harakahdaily.net	12	0.92%
ulamasedunia.org	11	0.84%
co.cc	11	0.84%
urusniaga.my	9	0.69%

Top 10 TLDs

Domain	Docs	% of total
com.my	667	51.07%
com	475	36.37%
org	59	4.52%
my	30	2.30%
net	25	1.91%
cc	11	0.84%
moe	7	0.54%
org.my	6	0.46%
eu	4	0.31%
edu.my	4	0.31%

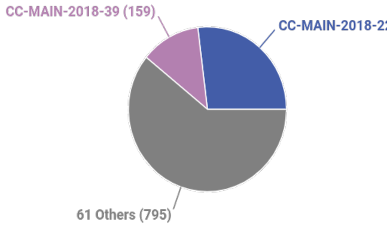
Documents size (in segments) ⓘ

≤ 25 segments **84.69%** (1.1K documents)
> 25 segments **15.31%** (200 documents)



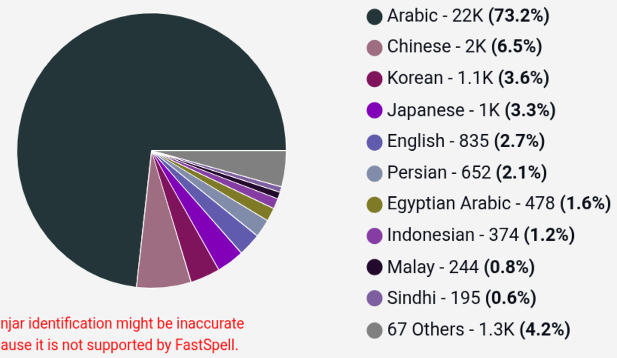
Document collections

CC = 91.58%
IA = 8.42%



Language Distribution

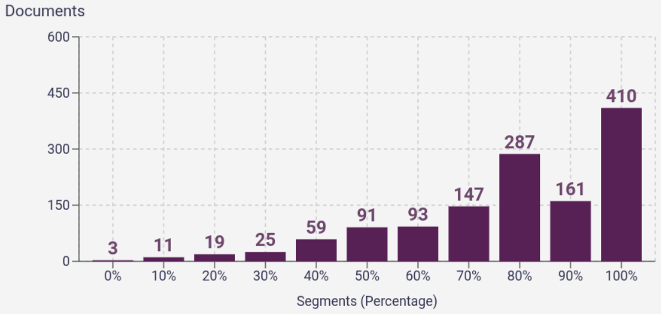
Number of segments in the Banjar corpus



*Banjar identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Banjar inside documents

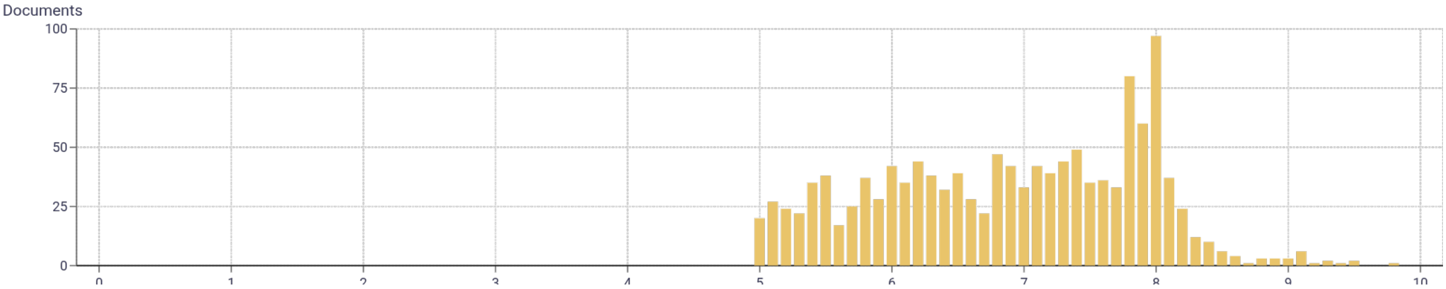
segments < 50% - **8.96%** (117 documents)
segments ≥ 50% - **91.04%** (1.2K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

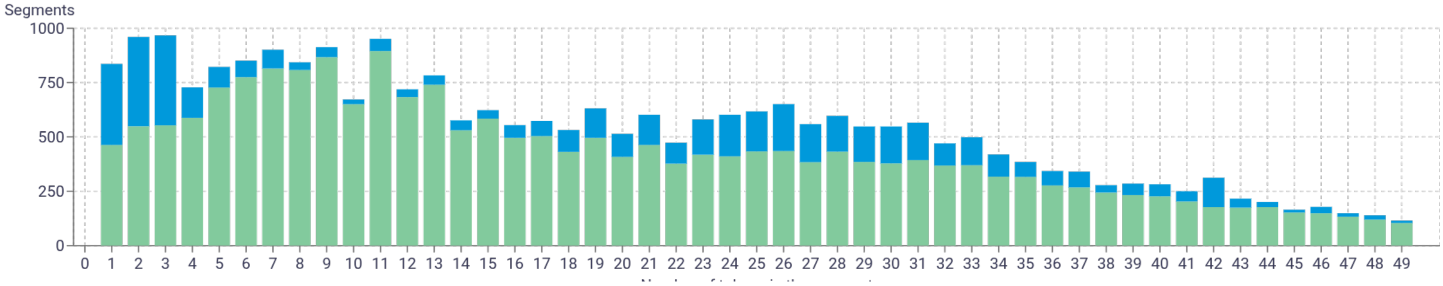
score ≥ 5 - **100%** (1.3K documents)



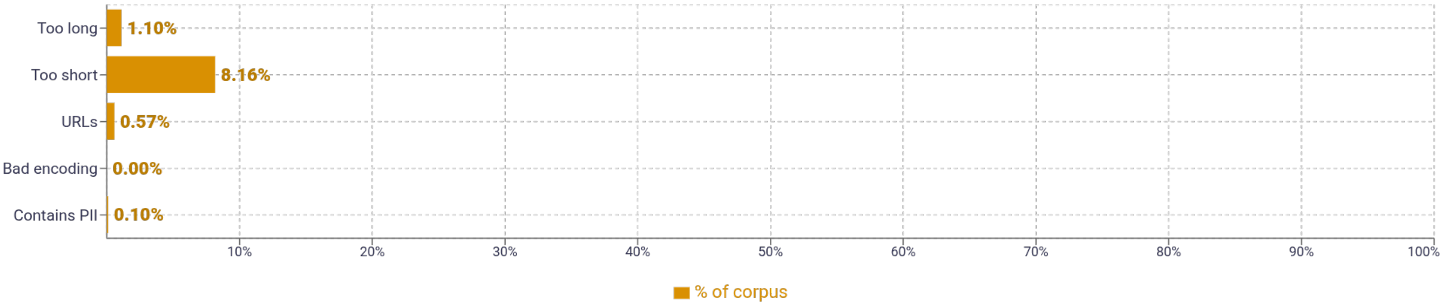
Segment length distribution by token

≤ **49** tokens = **26K** segments | **5.3K** duplicates

> **50** tokens = **3.9K** segments | **298** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	16,988 | يغ 8,255 | م 7,869 | ن 7,212 | و 6,597 | الله
2	arabic script | 4,102 yue chinese | 998 644 | رسول الله 590 | هاري اين 564 | اورغ يغ
3	464 | صلى الله عليه 463 | الله عليه وسلم 399 | الله سبحانه وتعالى 183 | الله صلى الله 182 | رسول الله صلى
4	448 | صلى الله عليه وسلم 179 | رسول الله صلى الله 179 | الله صلى الله عليه 123 | فنديديقن اسلم تيغكاتن ساتو 123 | سوكاتن بارو كريكولوم برسفادو
5	175 | رسول الله صلى الله عليه 170 | الله صلى الله عليه وسلم 123 | رنخغن تاهونن فنديديقن اسلم تيغكاتن 123 | تاهونن فنديديقن اسلم تيغكاتن ساتو 112 | محمد صلى الله عليه وسلم

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				