HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bjn_Latn	9/17/2025	Banjar (bjn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
21,227	363,963	284,756 (78.24 %)	12M	66,693,924	64.32 MB

Top 10 domains

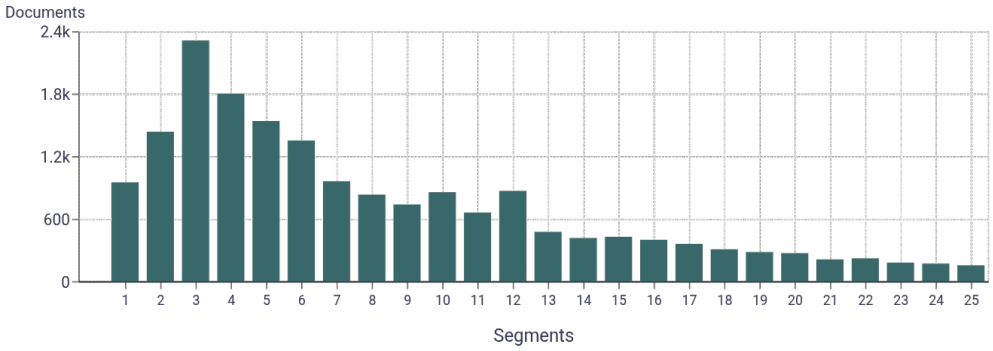
Domain	Docs	% of total
wikipedia.org	2.7K	12.48%
wordpress.com	1.1K	5.23%
blogspot.com	988	4.65%
tribunnews.com	640	3.02%
ruangguru.com	333	1.57%
banjarmasinbung...	302	1.42%
uinsby.ac.id	258	1.22%
web.app	255	1.20%
kanalkalimantan...	212	1.00%
d5d.org	183	0.86%

Top 10 TLDs

Domain	Docs	% of total
com	10K	47.89%
org	3.6K	17.18%
ac.id	2.2K	10.19%
co.id	731	3.44%
id	676	3.18%
net	514	2.42%
info	494	2.33%
go.id	482	2.27%
app	291	1.37%
co	215	1.01%

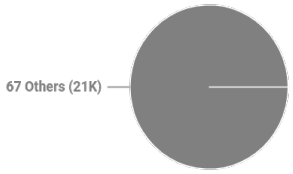
Documents size (in segments)

≤ 25 segments **86.23%** (18K documents)
> 25 segments **13.77%** (2.9K documents)



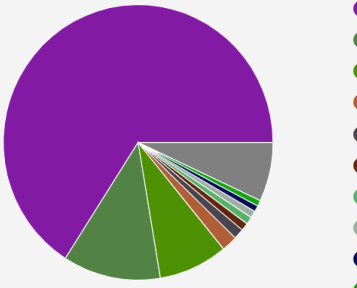
Document collections

CC = 91.24%
IA = 8.76%



Language Distribution

Number of segments in the Banjar (bjn) corpus

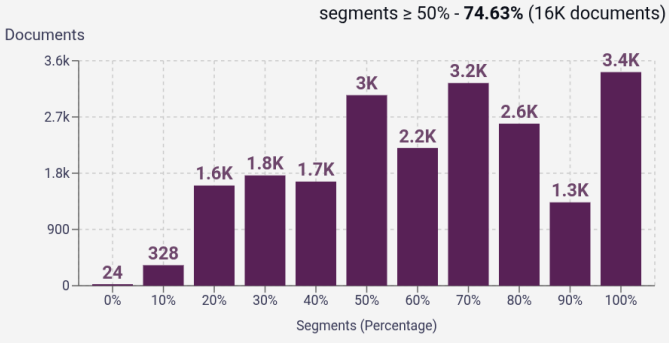


- Indonesian (id) - 240K **(65.9%)**
- English (en) - 43K **(11.7%)**
- Malay (ms) - 30K **(8.3%)**
- Italian (it) - 6.9K **(1.9%)**
- French (fr) - 4.5K **(1.2%)**
- Filipino (tl) - 3.3K **(0.9%)**
- Dutch (nl) - 3.2K **(0.9%)**
- German (de) - 3K **(0.8%)**
- Swedish (sv) - 2.6K **(0.7%)**
- Spanish (es) - 2.5K **(0.7%)**
- 137 Others - 25K **(6.9%)**

*Banjar (bjn) identification might be inaccurate because it is not supported by FastSpell

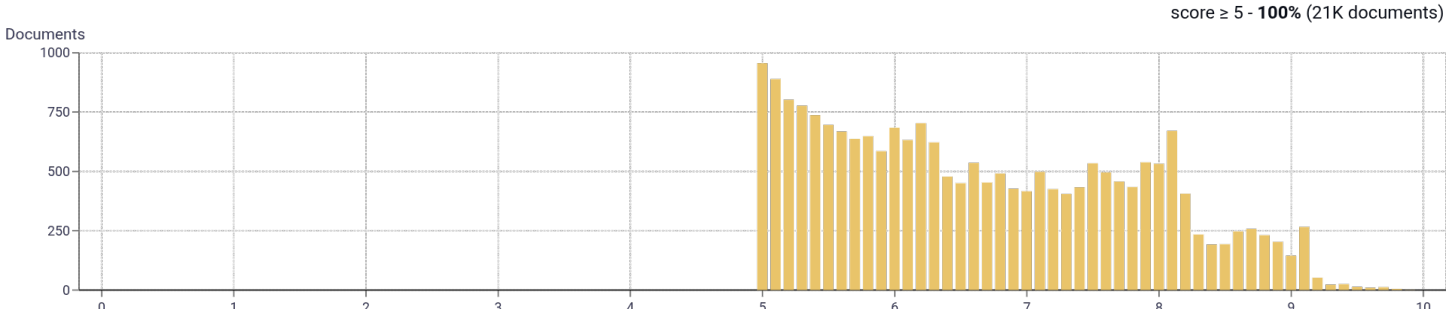
Percentage of segments in Banjar (bjn) inside documents

segments < 50% - **25.37%** (5.4K documents)
segments ≥ 50% - **74.63%** (16K documents)



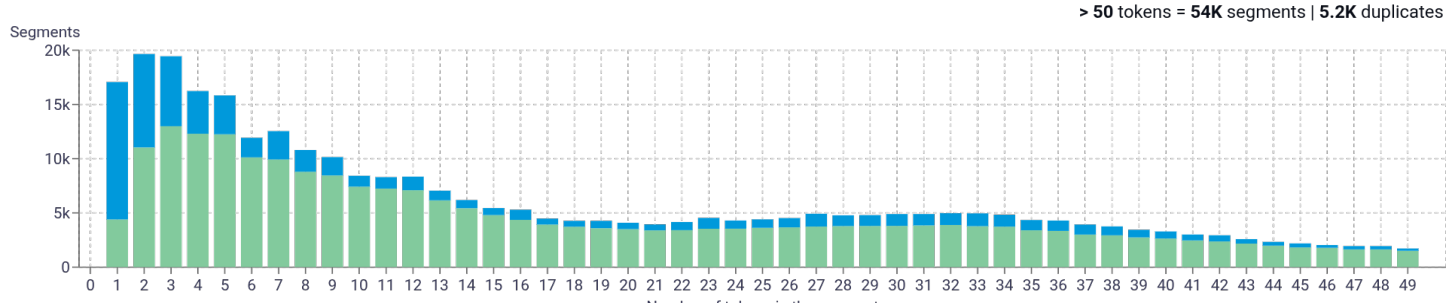
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (21K documents)

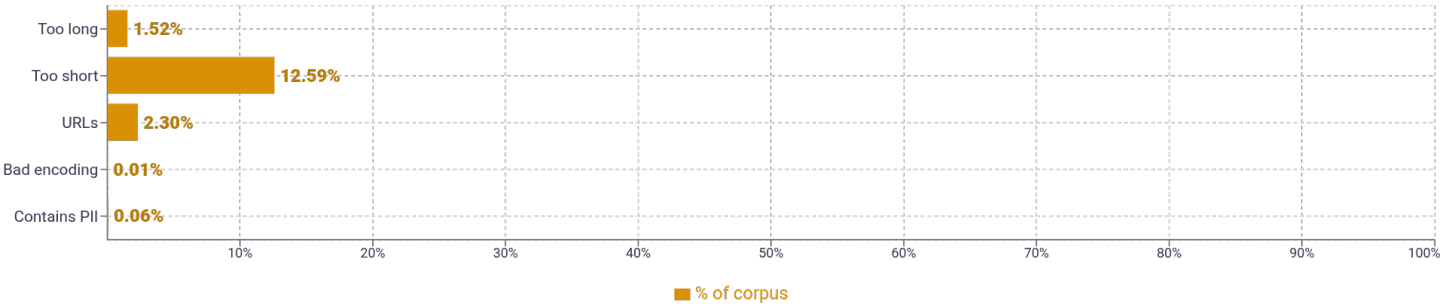


Segment length distribution by token

≤ 49 tokens = **310K** segments | **74K** duplicates
> 50 tokens = **54K** segments | **5.2K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	syair | 57,021 · ban | 36,765 · matriks | 32,320 · surabaya | 31,025 · sunan | 30,058
2	uin sunan | 25,448 · sunan ampel | 22,698 · ampel surabaya | 21,698 · undergraduate thesis | 20,573 · syair hk | 8,186
3	sunan ampel surabaya | 21,674 · uin sunan ampel | 19,931 · uin sunan kalijaga | 5,360 · sunan kalijaga yogyakarta | 4,212 · jarak mill laut | 2,232
4	uin sunan ampel surabaya | 19,605 · uin sunan kalijaga yogyakarta | 4,023 · iain sunan ampel surabaya | 1,686 · fc liverpool fc liverpool | 1,317 · liverpool fc liverpool fc | 1,306
5	liverpool fc liverpool fc liverpool | 1,295 · fc liverpool fc liverpool fc | 1,203 · mill jarak dari banjarmasin to | 1,173 · berapa mill jarak dari banjarmasin | 1,141 · jarak dari banjarmasin to halmahera | 1,003

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				