HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ban_Latn	9/16/2025	Balinese (ban)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
16,000	1,020,865	286,618 (28.08 %)	21M	113,835,703	109.92 MB

Top 10 domains

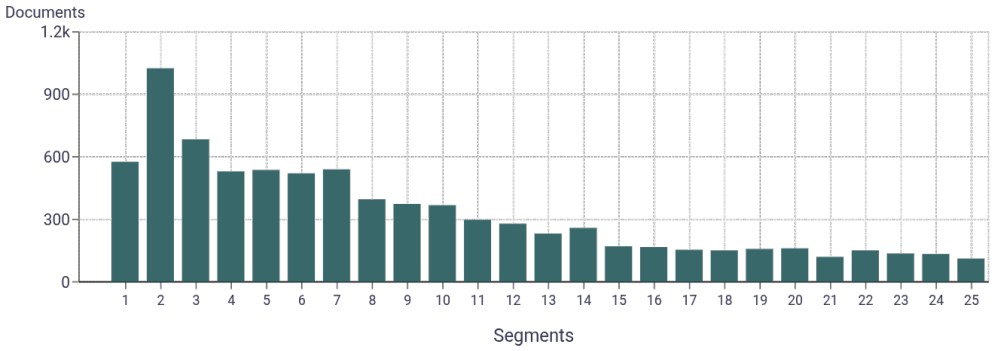
Domain	Docs	% of total
basabali.org	7.9K	49.56%
wikipedia.org	3K	18.90%
blogspot.com	871	5.44%
suarasakingbali...	487	3.04%
wordpress.com	347	2.17%
alkitab.mobi	307	1.92%
bible.is	262	1.64%
wikisource.org	260	1.63%
balitopnews.com	181	1.13%
blogspot.co.id	131	0.82%

Top 10 TLDs

Domain	Docs	% of total
org	11K	71.28%
com	3K	18.74%
mobi	307	1.92%
is	262	1.64%
co.id	150	0.94%
ac.id	120	0.75%
xyz	117	0.73%
net	86	0.54%
tv	82	0.51%
desa.id	77	0.48%

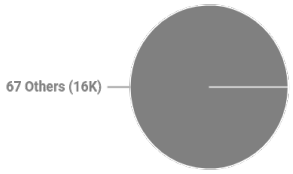
Documents size (in segments)

≤ 25 segments **51.46%** (8.2K documents)
> 25 segments **48.54%** (7.8K documents)



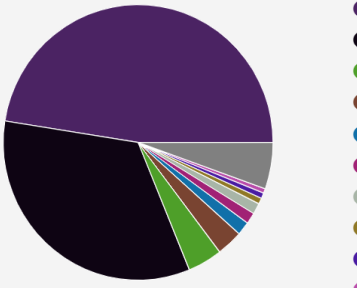
Document collections

CC = 96.14%
IA = 3.86%



Language Distribution

Number of segments in the Balinese (ban) corpus

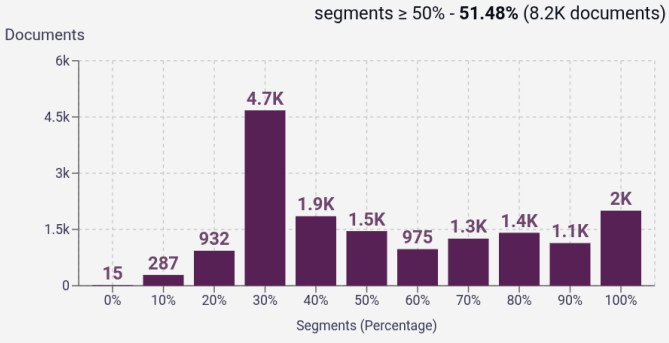


- English (en) - 485K **(47.5%)**
- Indonesian (id) - 344K **(33.7%)**
- Javanese (jv) - 42K **(4.1%)**
- Malay (ms) - 32K **(3.1%)**
- Filipino (tl) - 16K **(1.6%)**
- Italian (it) - 13K **(1.3%)**
- Hungarian (hu) - 13K **(1.3%)**
- Sundanese (su) - 7.3K **(0.7%)**
- Cebuano (ceb) - 6.7K **(0.7%)**
- Spanish (es) - 5.3K **(0.5%)**
- 134 Others - 56K **(5.5%)**

*Balinese (ban) identification might be inaccurate because it is not supported by FastSpell

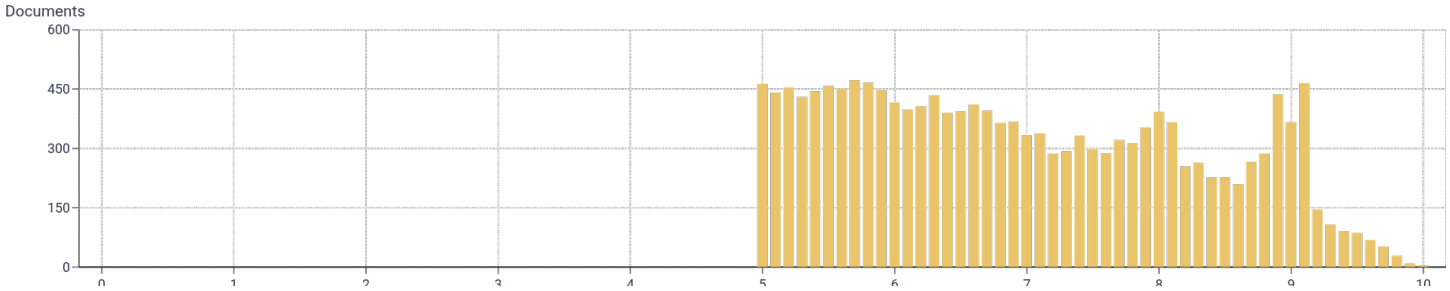
Percentage of segments in Balinese (ban) inside documents

segments < 50% - **48.52%** (7.8K documents)
segments ≥ 50% - **51.48%** (8.2K documents)



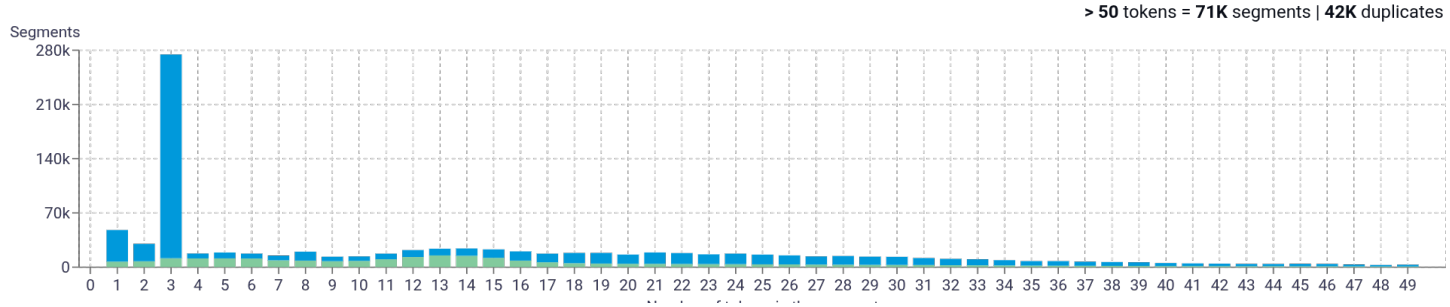
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (16K documents)

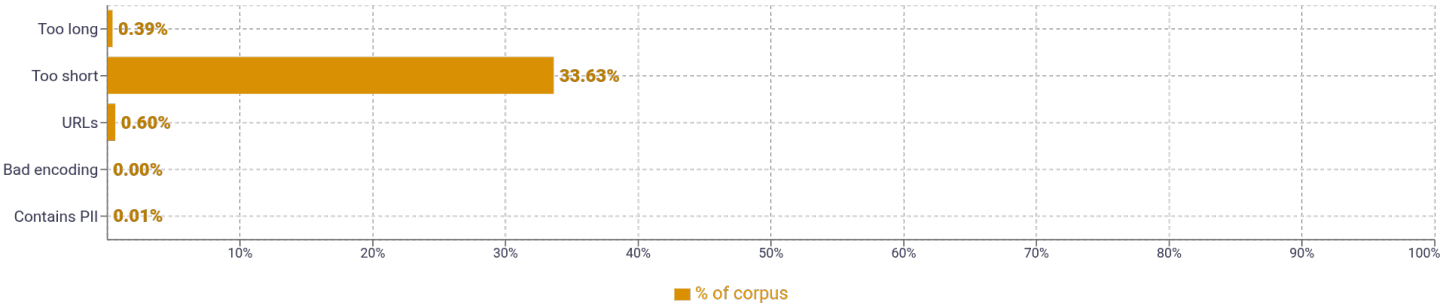


Segment length distribution by token

≤ 49 tokens = **950K** segments | **695K** duplicates
> 50 tokens = **71K** segments | **42K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	in | 685,052 · balinese | 233,260 · indonesian | 222,100 · english | 221,886 · bali | 161,933
2	in balinese | 216,969 · in indonesian | 216,310 · in english | 214,028 · of the | 11,838 · usage examples | 11,491
3	hyang widi wasa | 7,013 · usage examples pulled | 5,798 · pulled from the | 5,798 · examples pulled from | 5,798 · hyang widhi wasa | 3,874
4	usage examples pulled from | 5,798 · examples pulled from the | 5,798 · pulled from the virtual | 3,593 · from the virtual library | 3,593 · pulled from the community | 2,205
5	usage examples pulled from the | 5,798 · pulled from the virtual library | 3,593 · examples pulled from the virtual | 3,593 · pulled from the community spaces | 2,205 · examples pulled from the community | 2,205

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				