

General overview

Corpus	Date	Language
hplt-v3-ban_Latn	9/18/2025	Balinese

Volumes

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

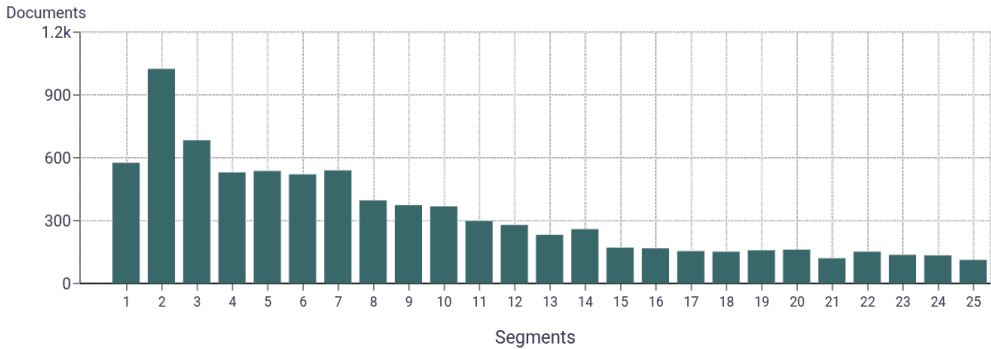
Top 10 domains

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

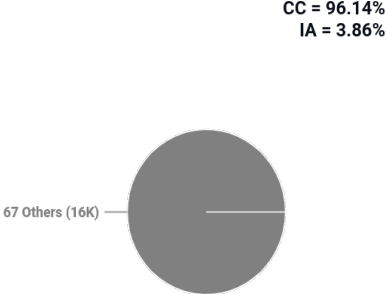
Domain	Docs	% of total
org	11K	71.28%
com	3K	18.74%
mobi	307	1.92%
is	262	1.64%
co.id	150	0.94%
ac.id	120	0.75%
xyz	117	0.73%
net	86	0.54%
tv	82	0.51%
desa.id	77	0.48%

Documents size (in segments) ⓘ



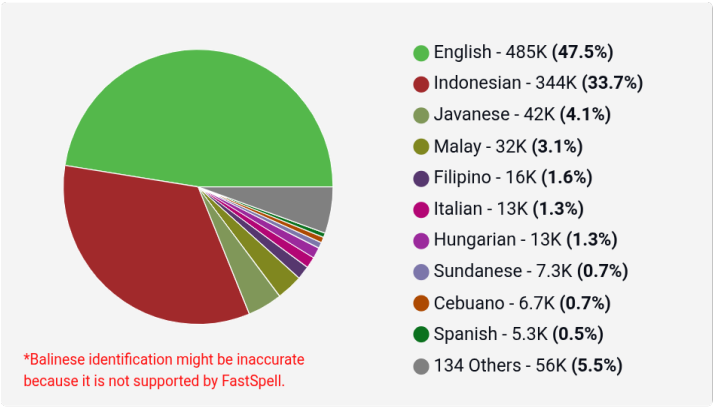
≤ 25 segments **51.46%** (8.2K documents)
> 25 segments **48.54%** (7.8K documents)

Document collections

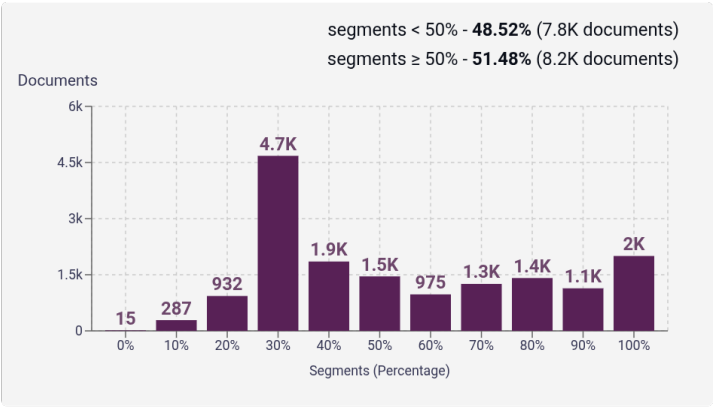


Language Distribution

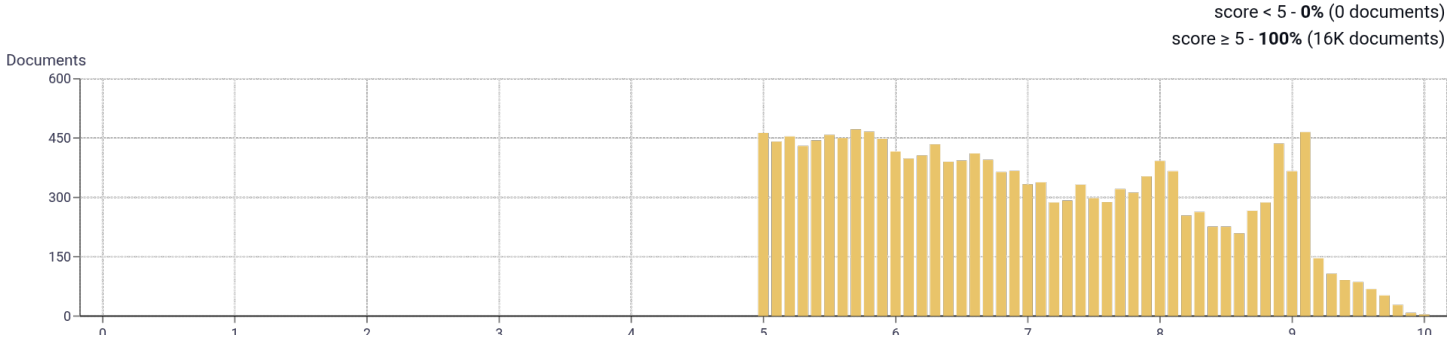
Number of segments in the Balinese corpus



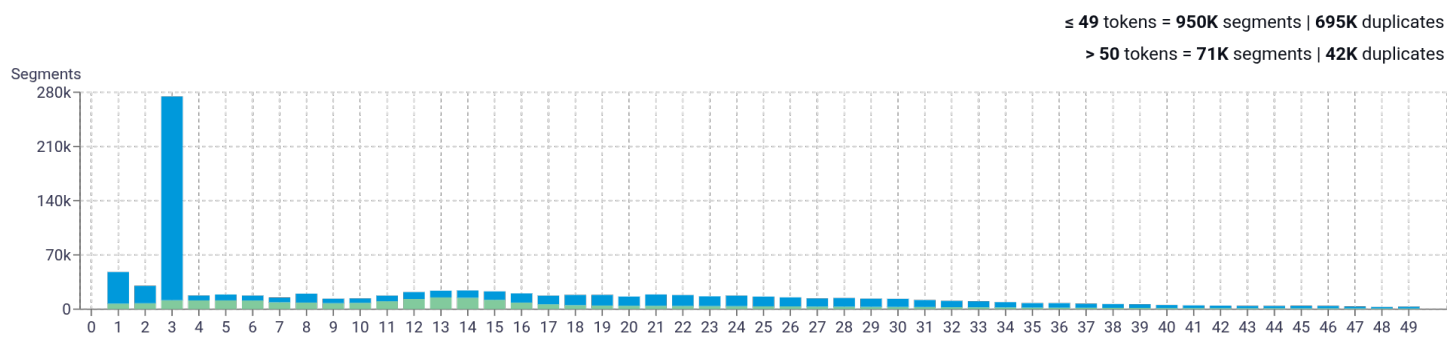
Percentage of segments in Balinese inside documents



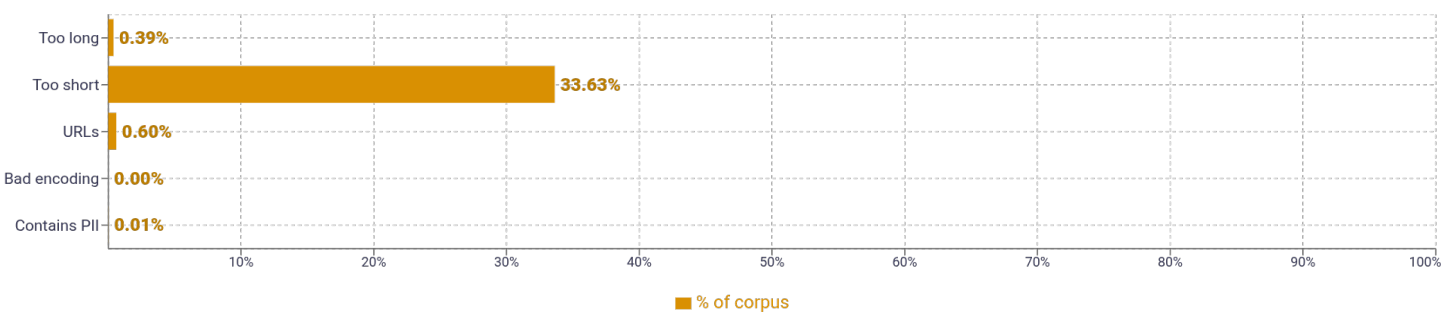
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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