HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bug_Latn	9/17/2025	Buginese

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,173	32,290	28,275 (87.57 %)	1.7M	8,599,225	8.79 MB

Top 10 domains

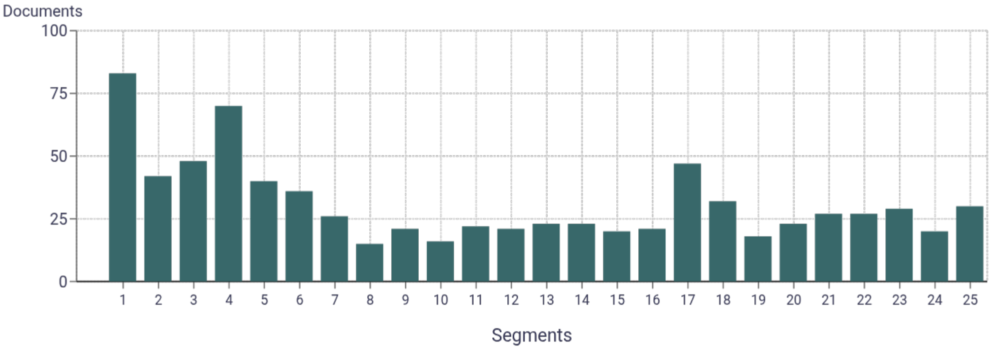
Domain	Docs	% of total
alkitab.mobi	535	45.61%
wordpress.com	110	9.38%
bible.is	46	3.92%
blogspot.com	37	3.15%
wikipedia.org	24	2.05%
indonesiachord.com	21	1.79%
chordpass.com	19	1.62%
basasulselwiki.org	17	1.45%
wikimedia.org	15	1.28%
ebible.org	15	1.28%

Top 10 TLDs

Domain	Docs	% of total
mobi	535	45.61%
com	369	31.46%
org	88	7.50%
is	46	3.92%
net	36	3.07%
id	25	2.13%
info	10	0.85%
pw	6	0.51%
de	6	0.51%
me	5	0.43%

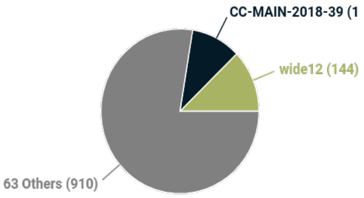
Documents size (in segments) ⓘ

≤ 25 segments **66.5%** (780 documents)
> 25 segments **33.5%** (393 documents)



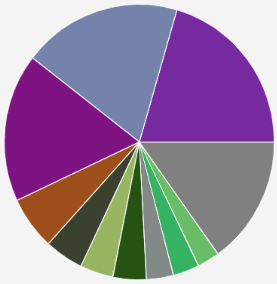
Document collections

CC = 82.01%
IA = 17.99%



Language Distribution

Number of segments in the Buginese corpus

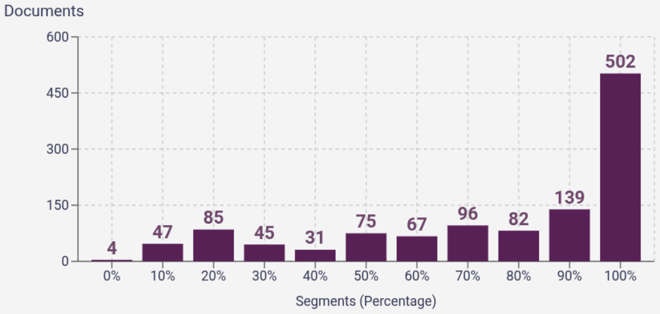


- ● Indonesian - 6.6K **(20.5%)**
- ● Sundanese - 6.1K **(19.0%)**
- ● English - 5.6K **(17.4%)**
- ● Malay - 2K **(6.3%)**
- ● French - 1.5K **(4.6%)**
- ● Hungarian - 1.3K **(4.0%)**
- ● Filipino - 1.3K **(4.0%)**
- ● Urdu - 1K **(3.2%)**
- ● Hindi - 1K **(3.1%)**
- ● German - 865 **(2.7%)**
- ● 95 Others - 4.9K **(15.1%)**

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

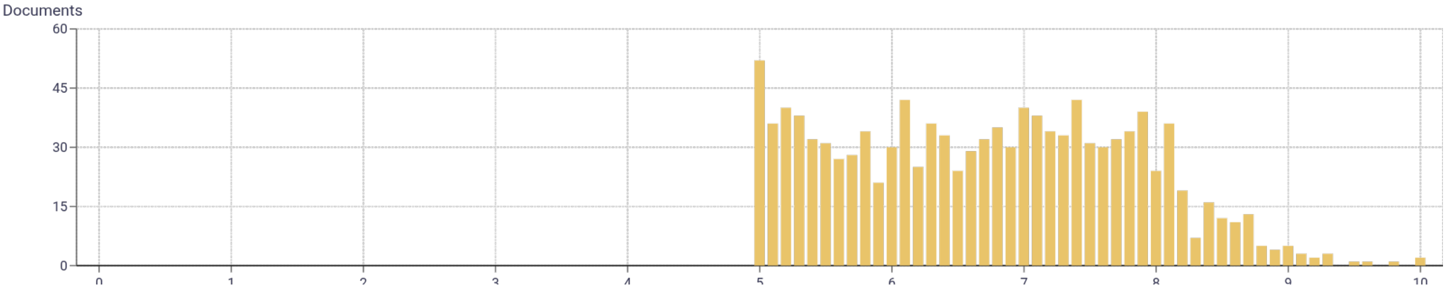
Percentage of segments in Buginese inside documents

segments < 50% - **18.07%** (212 documents)
segments ≥ 50% - **81.93%** (961 documents)



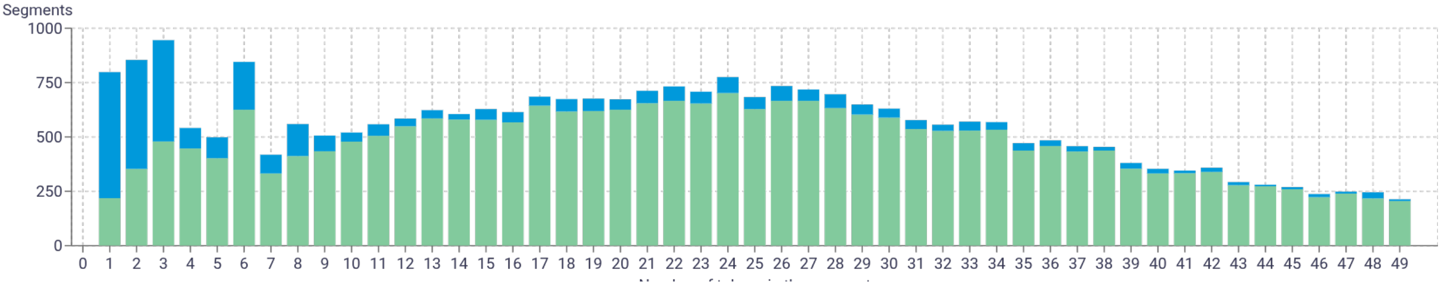
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.2K documents)

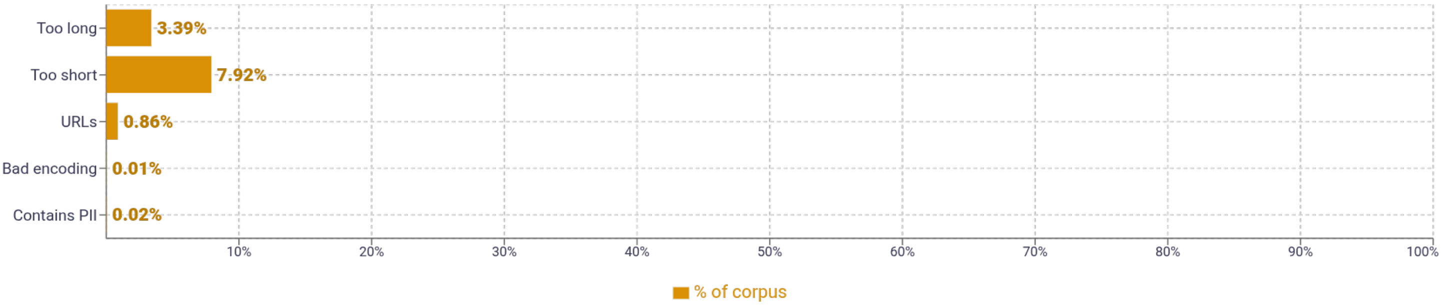


Segment length distribution by token

≤ 49 tokens = **27K** segments | **3.7K** duplicates
> 50 tokens = **4.9K** segments | **272** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	sibawa | 14,108 i | 8,080 lao | 8,014 anjo | 7,899 tpi | 7,082
2	anjo taua | 503 kamma anne | 497 trans tv | 426 sibawa sining | 388 iyaro sining | 376
3	lao ri mennang | 568 lao ri iko | 514 lao ri iyya | 327 mae ri allata | 295 lao ri puwangnge | 254
4	mae ri kau ngaseng | 173 puwangnge lao ri musa | 101 lalang ri kittaka angkanaya | 98 yèsus lao ri mennang | 88 yésus lao ri mennang | 84
5	nakkeda yèsus lao ri mennang | 48 name necklace with rhinestone letters | 35 custom name necklace with rhinestone | 35 moga saya tidak kena kutuk | 34 umma selleng malebbi engkae riamasei | 33

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				