HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-ace_Latn	9/16/2025	Achinese

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5,225	149,095	117,542 (78.84 %)	4.8M	25,253,142	25.01 MB

Top 10 domains

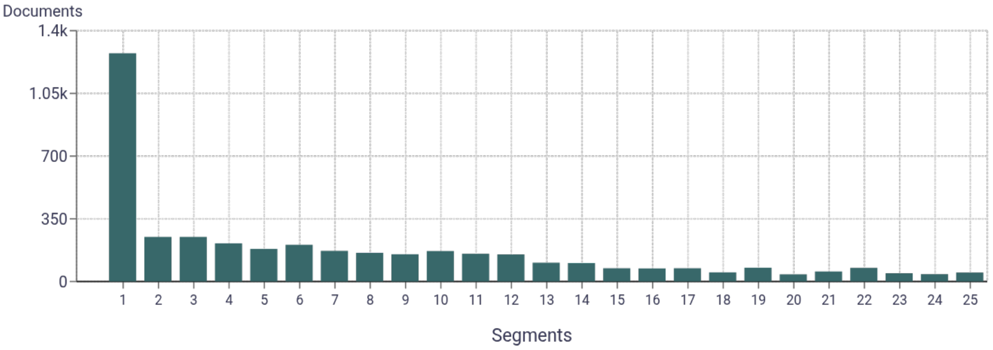
Domain	Docs	% of total
bible.is	1.2K	23.54%
wikipedia.org	669	12.80%
wordproject.org	444	8.50%
nasajaberita.com	288	5.51%
wordpress.com	201	3.85%
duhoctrungquoc.vn	147	2.81%
blogspot.com	122	2.33%
acehtrend.com	104	1.99%
petalokasi.org	73	1.40%
steemit.com	69	1.32%

Top 10 TLDs

Domain	Docs	% of total
com	1.8K	33.80%
org	1.4K	27.16%
is	1.2K	23.54%
vn	148	2.83%
net	83	1.59%
com.vn	61	1.17%
id	54	1.03%
cn	54	1.03%
co.id	52	1.00%
go.id	48	0.92%

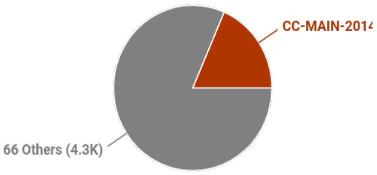
Documents size (in segments)

≤ 25 segments **80.33%** (4.2K documents)
> 25 segments **19.67%** (1K documents)



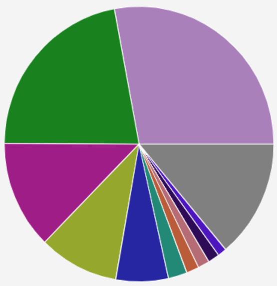
Document collections

CC = 92.90%
IA = 7.10%



Language Distribution

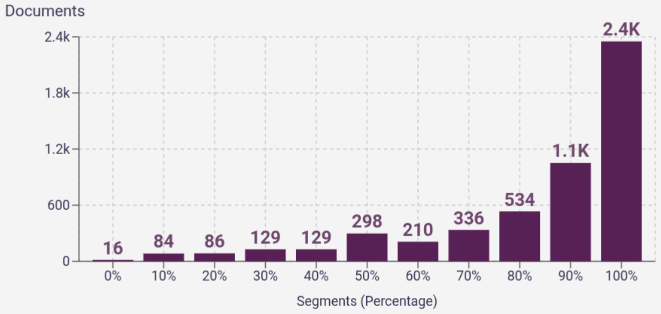
Number of segments in the Achinese corpus



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

Percentage of segments in Achinese inside documents

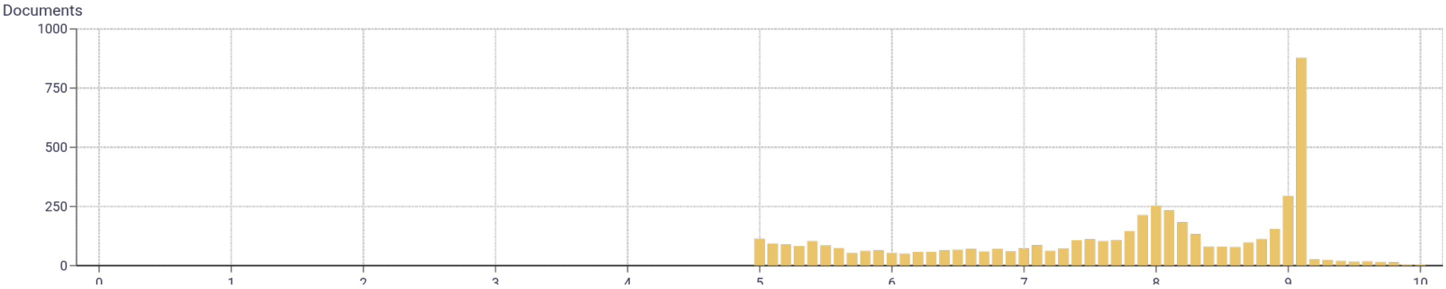
segments < 50% - **8.50%** (444 documents)
segments ≥ 50% - **91.50%** (4.8K documents)



Distribution of documents by document score

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

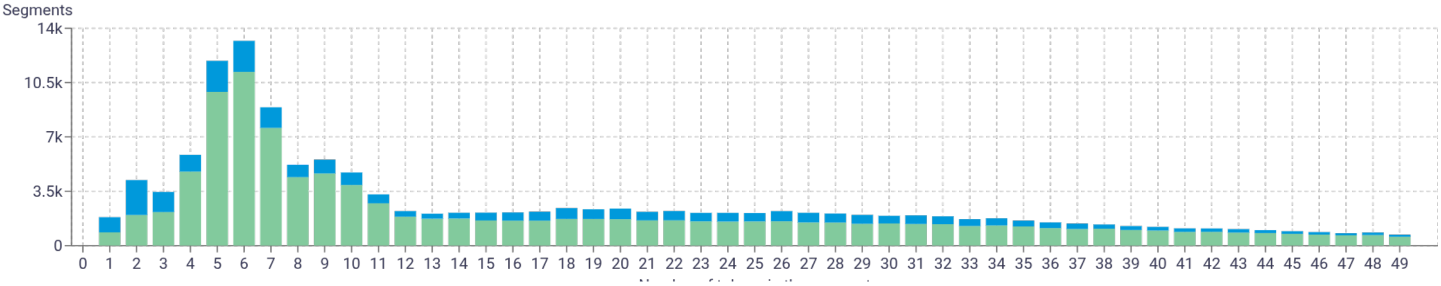
score ≥ 5 - **100%** (5.2K documents)



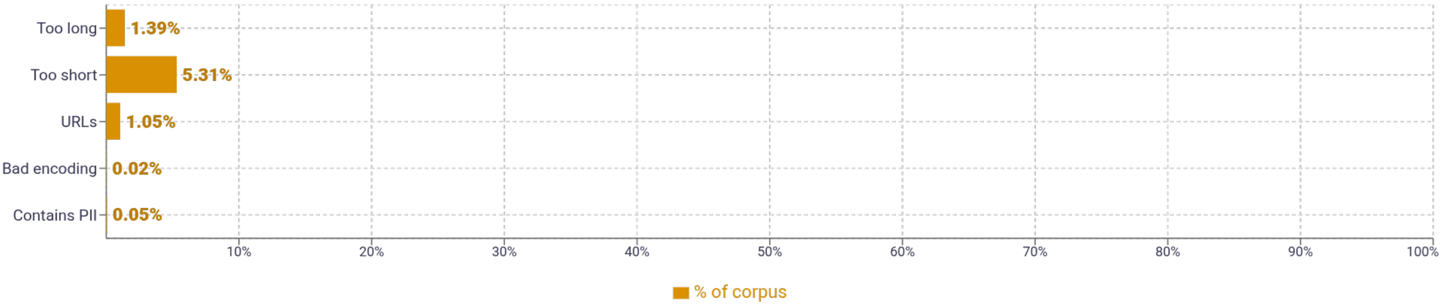
Segment length distribution by token

≤ 49 tokens = **134K** segments | **30K** duplicates

> 50 tokens = **15K** segments | **1.7K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	gata | 47,186 ureuëng | 45,266 ulôn | 39,361 kheueh | 38,816 bak | 32,840
2	teu allah | 6,468 latin script | 5,988 meunan cit | 5,808 nabi musa | 3,089 bak watée | 2,720
3	óh ka lheueh | 1,379 meunan cit deungon | 1,315 tuhan po teu | 1,152 ubak nabi musa | 909 teu allah gata | 879
4	tuhan po teu allah | 1,150 seulgi lee seulgi lee | 645 lé po teu allah | 642 ubak po teu allah | 637 lee seulgi lee seulgi | 631
5	tuhan po teu allah gata | 767 seulgi lee seulgi lee seulgi | 524 lee seulgi lee seulgi lee | 517 lé tuhan po teu allah | 246 tuhan meufeureuman ubak nabi musa | 178

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				