

General overview

Corpus	Date	Language
hplt-v3-kac_Latn	9/18/2025	Kachin

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
9,032	149,212	116,660 (78.18 %)	6.9M	28,381,179	27.51 MB

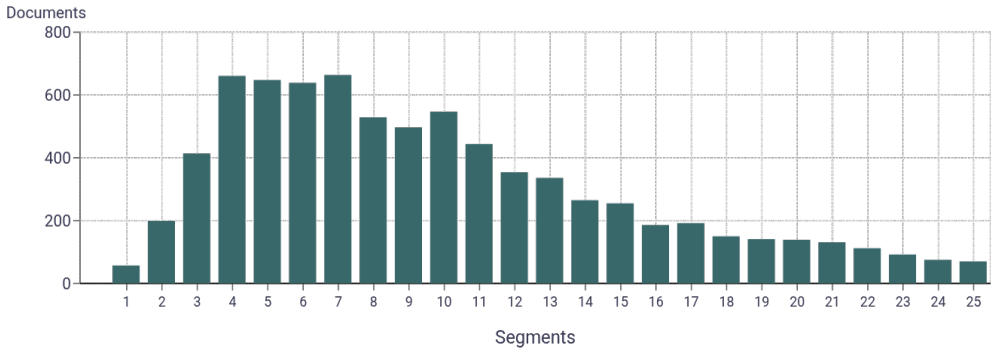
Top 10 domains

Domain	Docs	% of total
kachinnews.com	1.8K	19.46%
blogspot.com	1.7K	19.29%
rvasia.org	1.2K	13.73%
dehong.gov.cn	1.2K	13.08%
kachinnet.net	669	7.41%
kachinlandnews.com	348	3.85%
jw.org	315	3.49%
hkakaborazi.net	304	3.37%
myutsawmyit.net	138	1.53%
blogspot.sg	132	1.46%

Top 10 TLDs

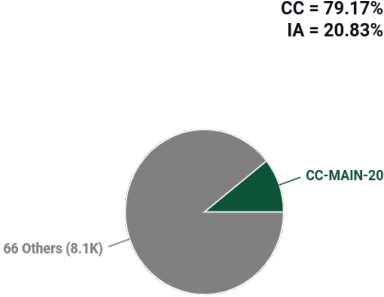
Domain	Docs	% of total
com	4.4K	48.17%
org	1.9K	21.32%
gov.cn	1.2K	13.08%
net	1.1K	12.42%
sg	132	1.46%
co.nz	119	1.32%
ca	50	0.55%
in	35	0.39%
edu.au	29	0.32%
org.sg	19	0.21%

Documents size (in segments) ⓘ



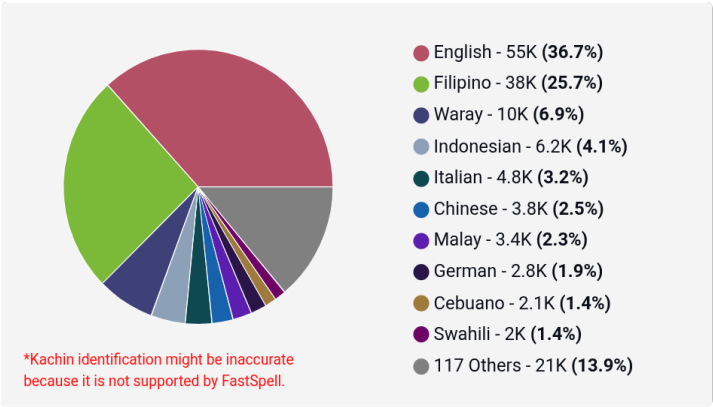
≤ 25 segments **86.33%** (7.8K documents)
> 25 segments **13.67%** (1.2K documents)

Document collections

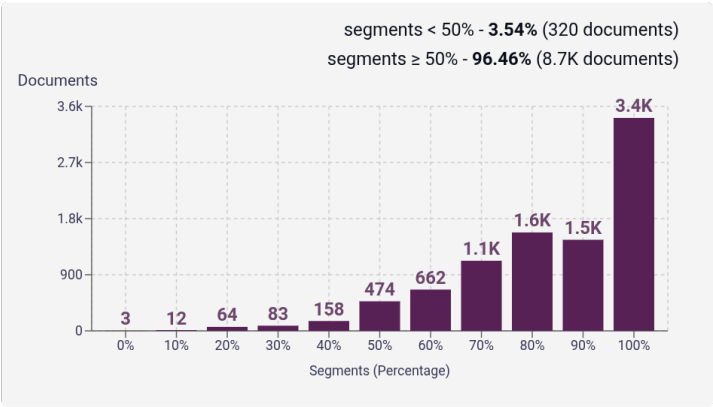


Language Distribution

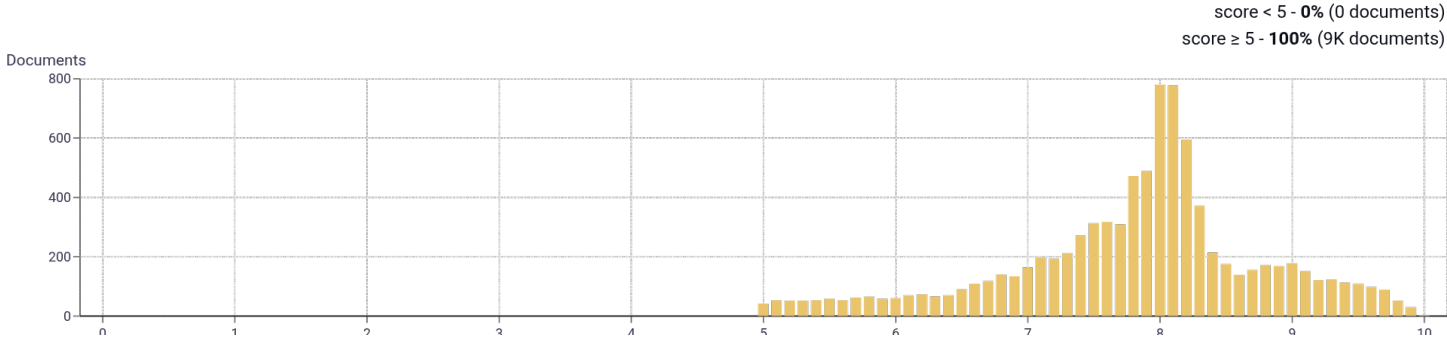
Number of segments in the Kachin corpus



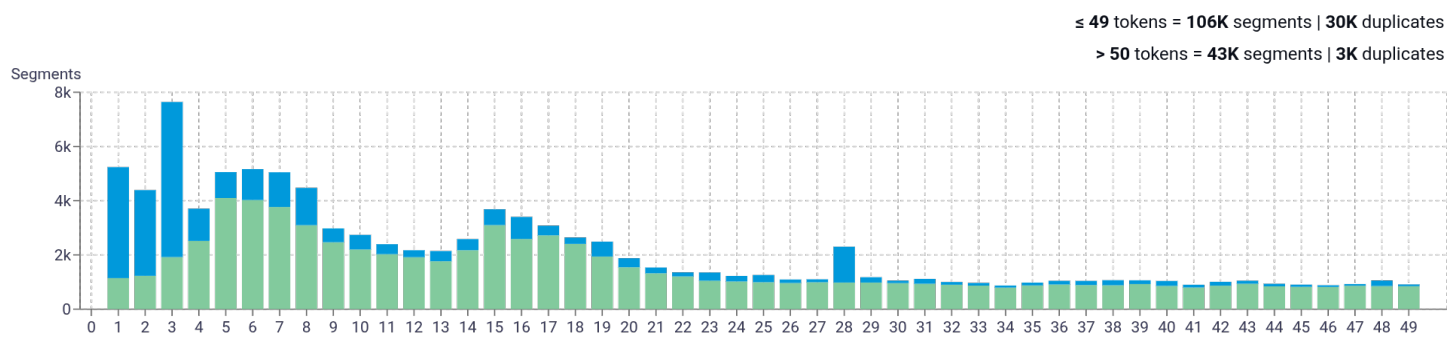
Percentage of segments in Kachin inside documents



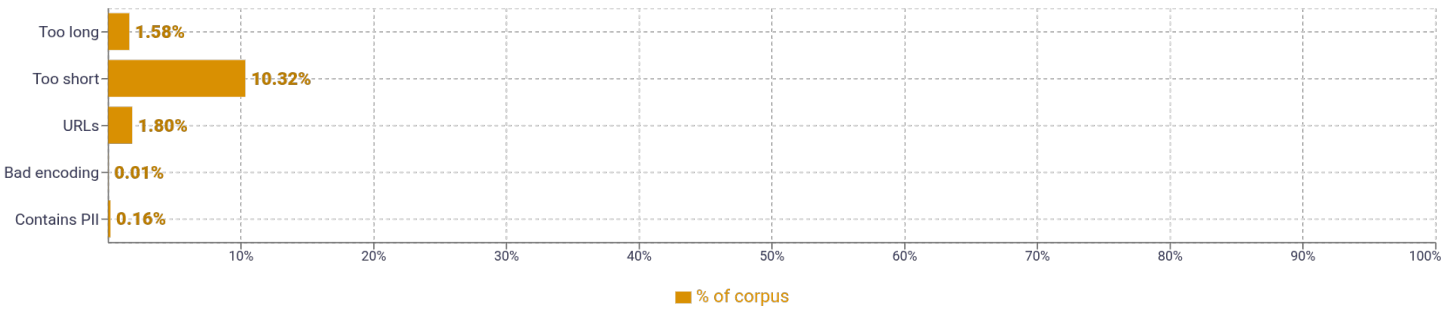
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ai 450,317 ni 161,989 hpe 140,949 nga 140,801 gaw 91,650	
2	nga ai 79,670 ai lam 44,449 ni hpe 30,599 rai nga 26,526 ra ai 16,604	
3	rai nga ai 22,518 chye lu ai 10,966 ai lam ni 7,698 nga ma ai 5,092 ai lam hpe 4,935	
4	lam chye lu ai 3,932 ai rai nga ai 2,419 ai lam ni hpe 2,119 lam rai nga ai 1,852 hpe chye lu ai 1,851	
5	ai lam chye lu ai 2,207 ai lam rai nga ai 1,481 radio veritas asia buick st 1,223 lam na chye lu ai 1,090 re lam chye lu ai 1,081	

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				