HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bod_Tibt	9/18/2025	Tibetan

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
27,863	480,612	396,286 (82.45 %)	122M	177,823,758	473.58 MB

Top 10 domains

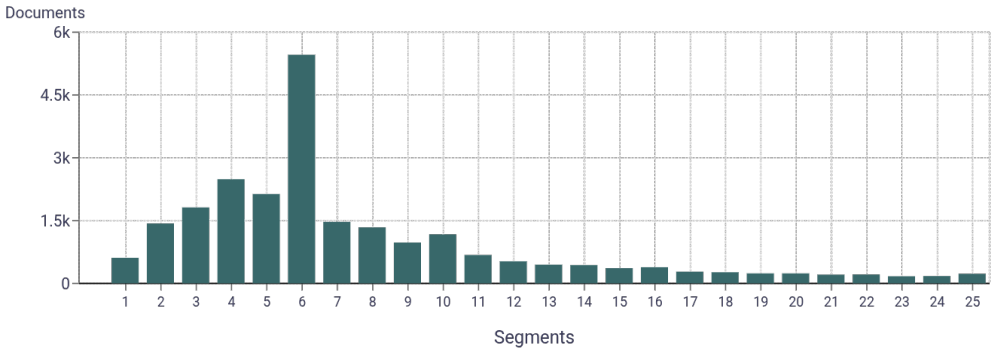
Domain	Docs	% of total
84000.co	5K	17.92%
tibettimes.net	2K	7.24%
bod.asia	1.9K	6.97%
tibet3.com	1.7K	6.05%
yongzin.com	1.4K	4.92%
zangdiyg.com	881	3.16%
people.com.cn	873	3.13%
tsadra.org	827	2.97%
tibet.cn	556	2.00%
chithu.org	519	1.86%

Top 10 TLDs

Domain	Docs	% of total
com	9.2K	33.13%
co	5K	17.92%
org	4.8K	17.28%
net	2.8K	9.95%
asia	2K	7.28%
cn	1.8K	6.45%
com.cn	918	3.29%
edu	369	1.32%
gov.cn	252	0.90%
us	74	0.27%

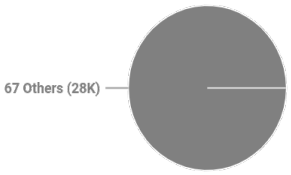
Documents size (in segments)

≤ 25 segments **85.3%** (24K documents)
> 25 segments **14.7%** (4.1K documents)



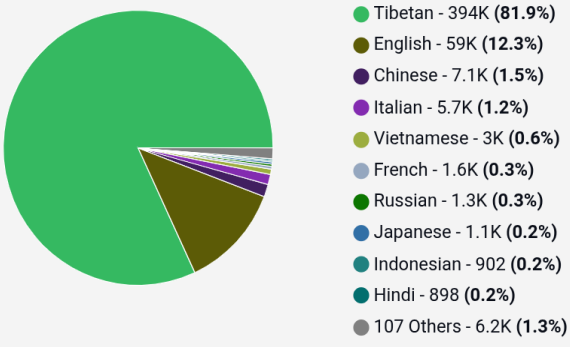
Document collections

CC = 95.21%
IA = 4.79%



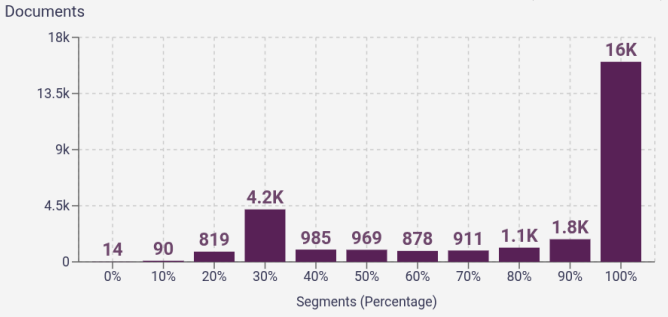
Language Distribution

Number of segments in the Tibetan corpus



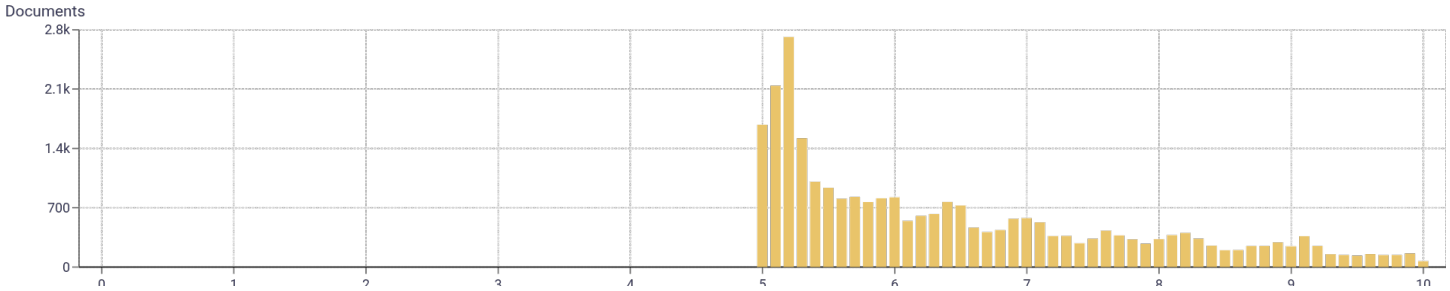
Percentage of segments in Tibetan inside documents

segments < 50% - **21.94%** (6.1K documents)
segments ≥ 50% - **78.06%** (22K documents)



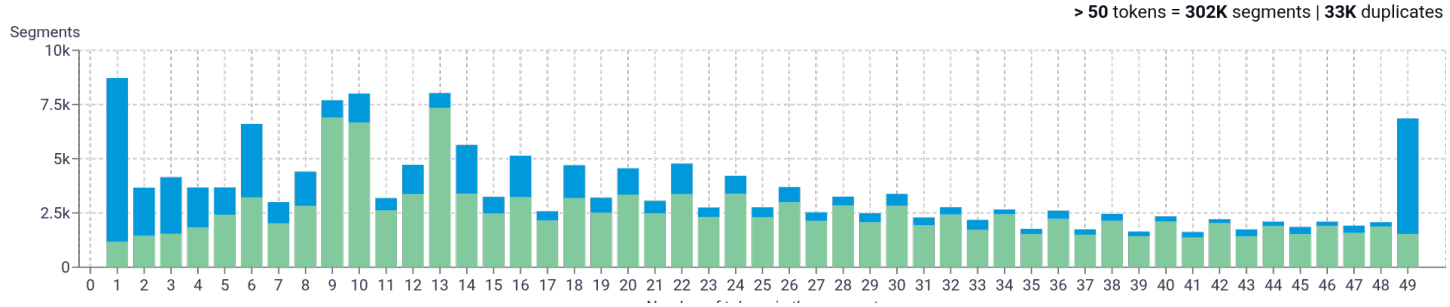
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (28K documents)

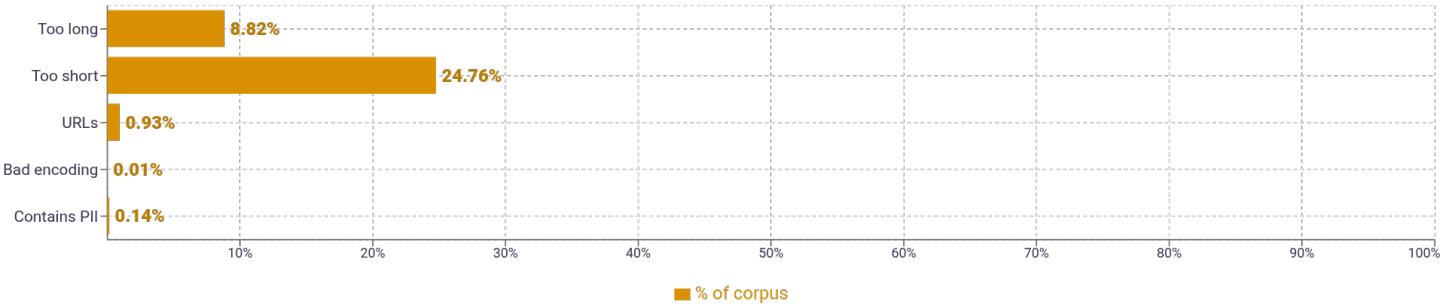


Segment length distribution by token

≤ 49 tokens = **178K** segments | **51K** duplicates
> 50 tokens = **302K** segments | **33K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	གས | 1,431,122 · ང་ | 986,110 · བས | 665,867 · པར | 611,565 · འད | 538,381
2	of the | 17,667 · the tibetan | 10,831 · the u | 9,990 · the ekangyur | 9,972 · in the | 9,075
3	the university of | 5,064 · university of virginia | 5,005 · tibetan and himalayan | 5,005 · the tibetan and | 5,001 · a variety of | 4,999
4	the tibetan and himalayan | 4,996 · the university of virginia | 4,992 · tibetan and himalayan library | 4,991 · used in this translation | 4,988 · translations are made from | 4,988
5	the tibetan and himalayan library | 4,990 · translations are made from a | 4,988 · tibetan page displayed here is | 4,988 · the tibetan page displayed here | 4,988 · the readings used in this | 4,988

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				