HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-dzo_Tibt	9/18/2025	Dzongkha

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
90	88,549	82,483 (93.15 %)	13M	19,806,489	51.11 MB

Top 10 domains

Domain	Docs	% of total
tipitakapali.org	77	85.56%
biblearc.com	6	6.67%
ebible.org	4	4.44%
wikipedia.org	1	1.11%
forgottenlangua...	1	1.11%
blogspot.com	1	1.11%

Top 10 TLDs

Domain	Docs	% of total
org	83	92.22%
com	7	7.78%

Documents size (in segments) ⓘ
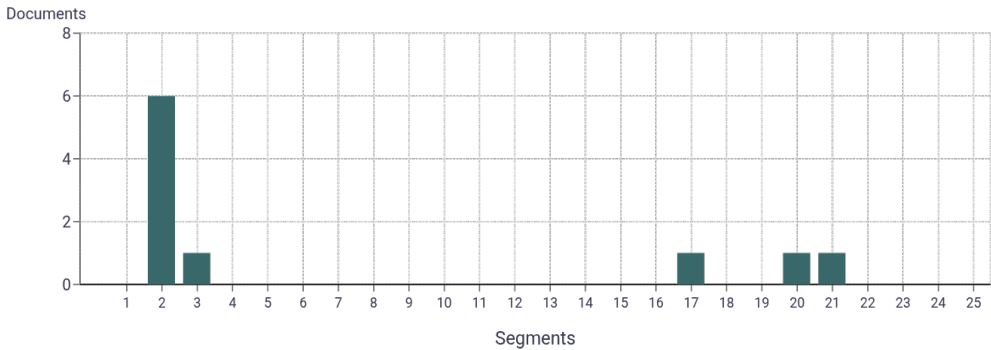
≤ 25 segments **11.11%** (10 documents)
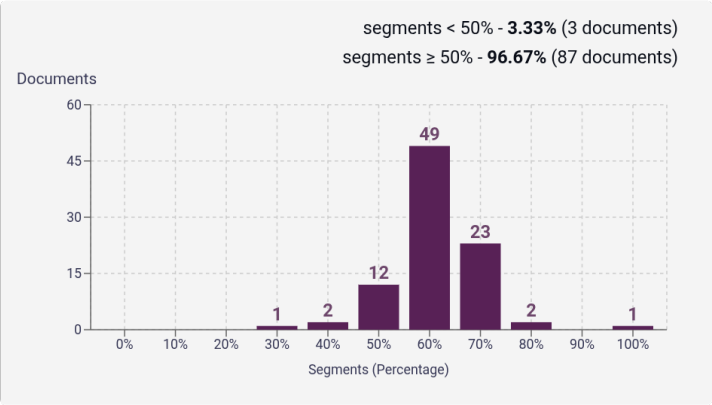> 25 segments **88.89%** (80 documents)

Document collections

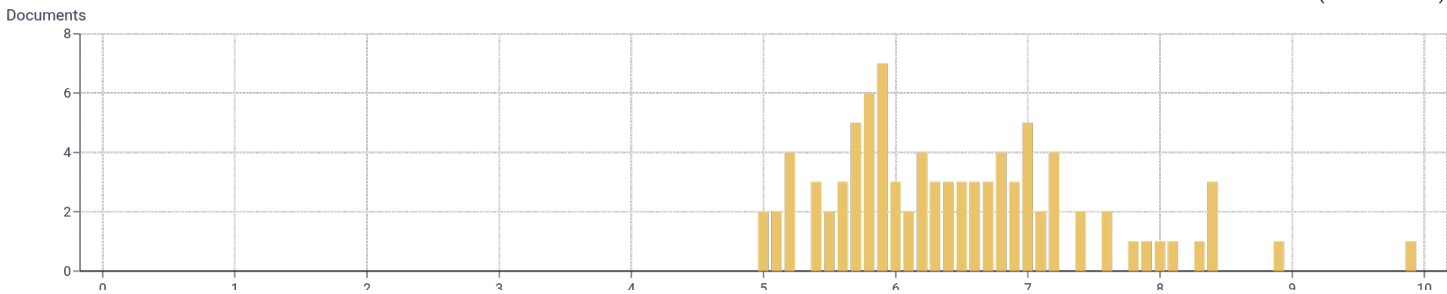CC = 100.00%
IA = 0.00%
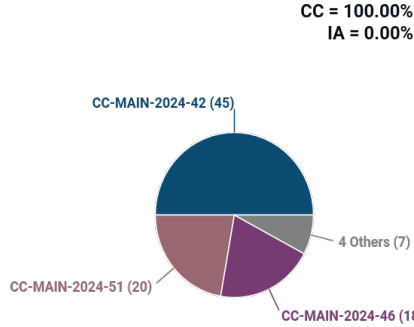
Language Distribution

Number of segments in the Dzongkha corpus

- Tibetan - 29K **(33.2%)**
- Russian - 26K **(29.7%)**
- Indonesian - 12K **(13.3%)**
- Chinese - 8.8K **(9.9%)**
- English - 4.3K **(4.9%)**
- Japanese - 2.1K **(2.3%)**
- Tamil - 961 **(1.1%)**
- Cebuano - 790 **(0.9%)**
- Spanish - 544 **(0.6%)**
- Serbian - 534 **(0.6%)**
- 63 Others - 3.1K **(3.5%)**

*Dzongkha identification might be inaccurate because it is not supported by FastSpell.
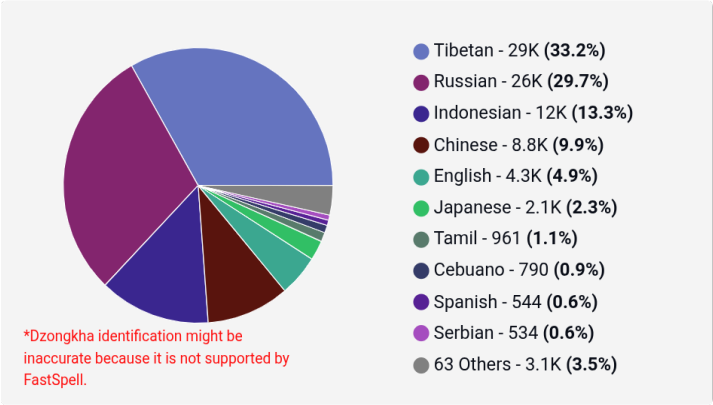
Percentage of segments in Dzongkha inside documents

segments < 50% - **3.33%** (3 documents)
segments ≥ 50% - **96.67%** (87 documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (90 documents)

Segment length distribution by token

≤ **49** tokens = **48K** segments | **6K** duplicates

> **50** tokens = **40K** segments | **59** duplicates

Segments

3.8k

2.85k

1.9k

950

0

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49

Segment noise distribution

Too long	**2.66%**
Too short	**15.60%**
URLs	**0.00%**
Bad encoding	**0.00%**
Contains PII	**0.00%**

10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

■ % of corpus
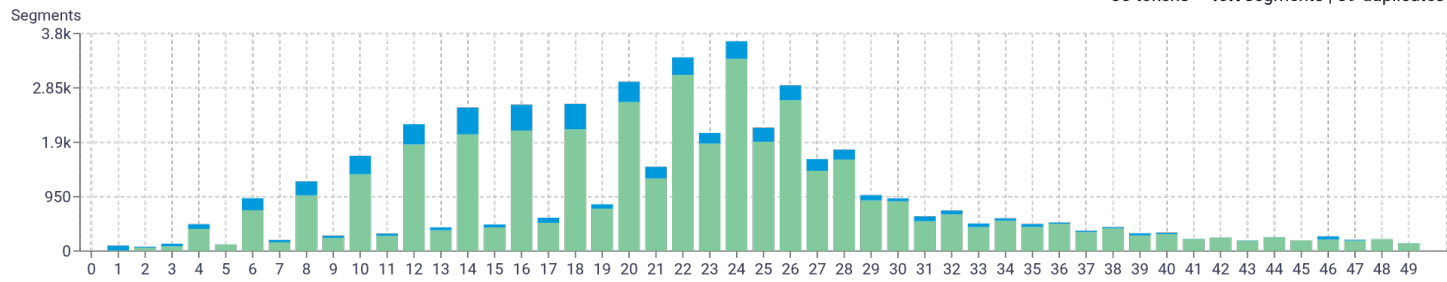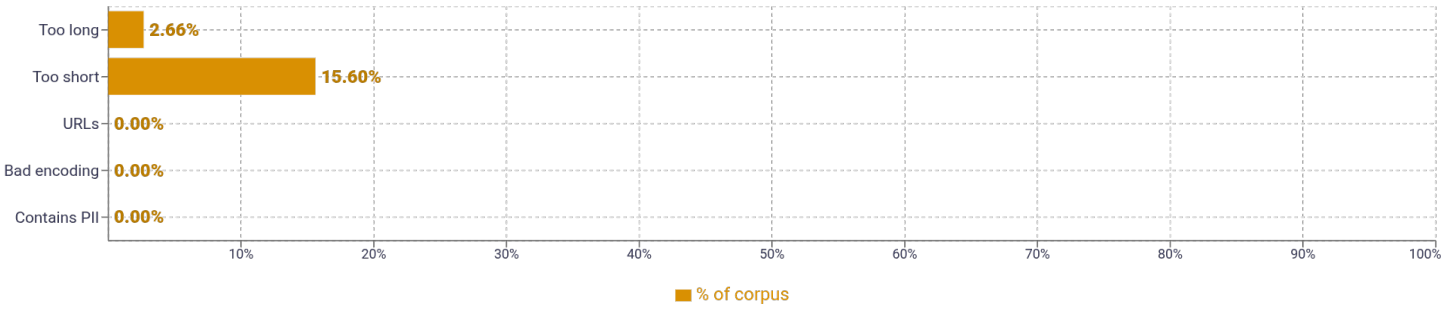
Frequent n-grams

SIZE	N-GRAMS					
1	ད	701,779	ན	562,749	ས	336,738	ད	306,031	ཨ	276,029	
2	ན ས	10,174	ན ཨ	9,437	ན ཚ	5,298	ན ད	4,781	ན ད	4,154	
3	ན ཚ ཨ	281	ན ཚ ས	198	ན ཚ ད	142	ན ཚ ན	132	ཡ ཚ ཨ	131	
4	ཚ ན ཚ ད	20	ཚ ན ཚ ཨ	10	ཨཐ ཚ ན པཉ	8	ན ཚ ན ཤཀ	8	ཚ ཚ ཛ ཛ	7	
5	ར ལ སྱ ས ད	5	ད ཚ ཚ ཛ ཛ	4	ཛ པཐ དྲ གར སཡ	3	ན པ ས བབ	3	ཙ ད ཐ ད ད	3	

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				