

General overview

Corpus	Analytics date	Language
dzo_Tibt.jsonl.tsv	11/27/2024	Dzongkha (dz)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,626	39,971	18,450 (46.16 %)	5M	19.8 MB	7,336,913

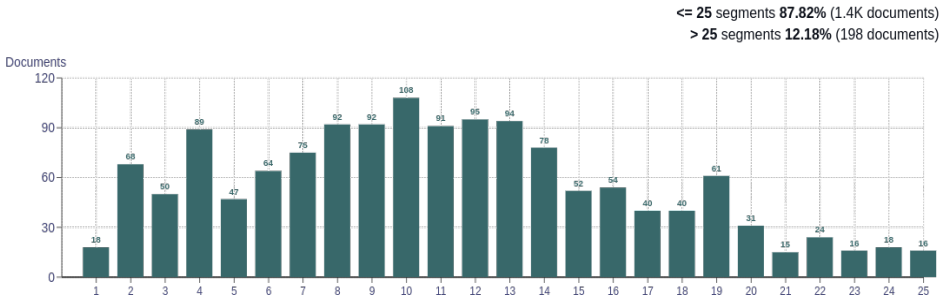
Top 10 domains

Domain	Docs	% of total
blogspot.com	477	29.34
dzkuensel.com	429	26.38
libreoffice.org	153	9.41
ecb.bt	83	5.10
gnhc.gov.bt	50	3.08
wikipedia.org	49	3.01
investigative-manual.org	45	2.77
virginia.edu	37	2.28
library.gov.bt	31	1.91
dzkuensel.bt	28	1.72

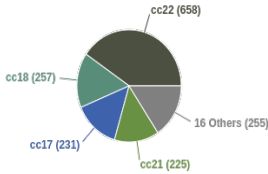
Top 10 TLDs

Domain	Docs	% of total
com	938	57.69
org	251	15.44
gov.bt	212	13.04
bt	138	8.49
edu	37	2.28
edu.bt	12	0.74
in	9	0.55
net	8	0.49
nl	7	0.43
de	7	0.43

Documents size (in segments)

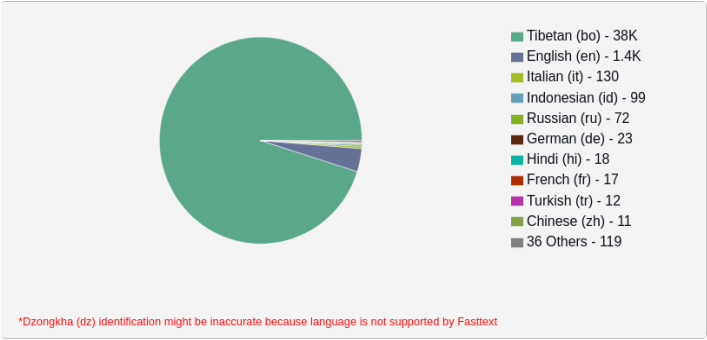


Documents by collection

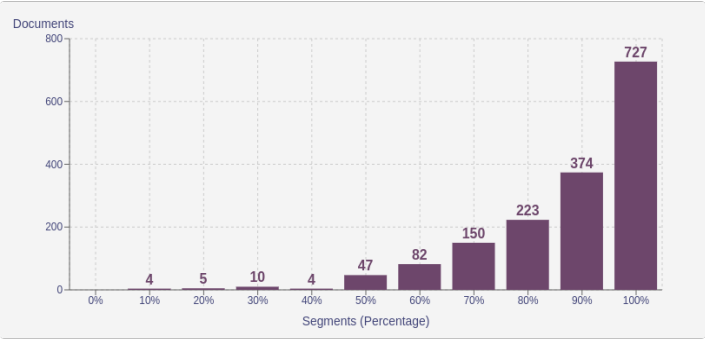


Language Distribution

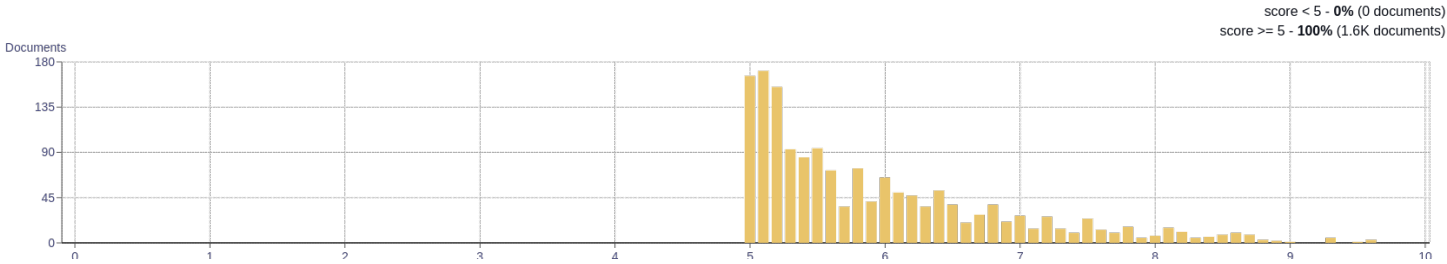
Number of segments



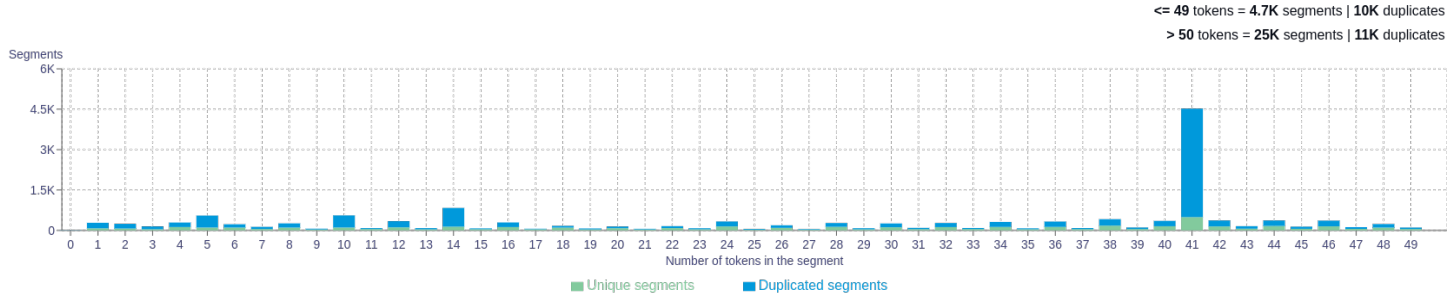
Percentage of segments in Dzongkha (dz) inside documents



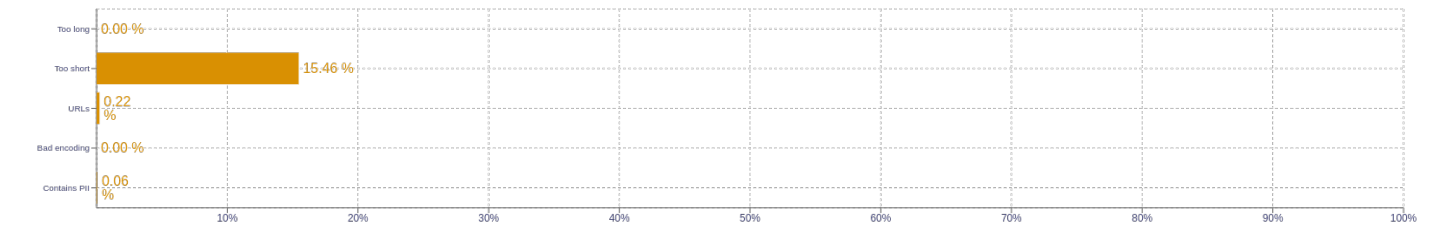
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ཁ 175207ཁྱ 167394ལ 143180ལྔ 128471ལྔ 78947</div>
2	<div>no comments 4349pm no 3731am no 618read more 192in the 153</div>
3	<div>pm no comments 3731am no comments 618dasho zoepoen wangchukཉ 66you want to 21posted in ཁྱ 20</div>
4	<div>text to speech synthesis 14pioneering dzongkha text to 14dzongkha text to speech 14that you want to 10you can enter a 8</div>
5	<div>pioneering dzongkha text to speech 14dzongkha text to speech synthesis 14use the date series type 8type and this option to 8the date series type and 8</div>

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 "number of 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>, , , 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>