HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-mya_Mymr	9/18/2025	Burmese (my)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,979,319	36,789,487	26,914,866 (73.16 %)	1.2B	7,183,215,104	18.5 GB

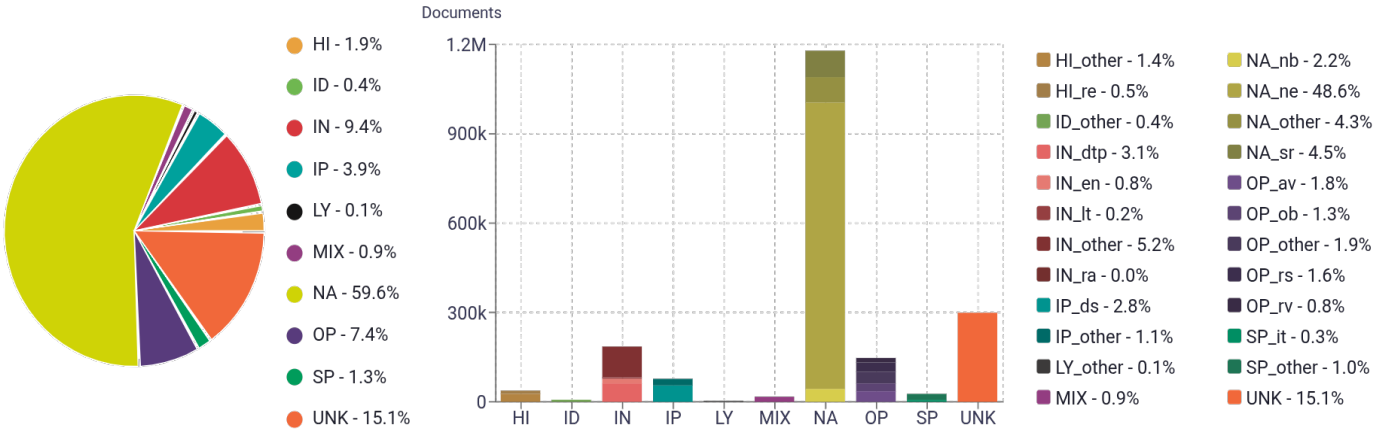
Top 10 domains

Domain	Docs	% of total
blogspot.com	111K	5.60%
voanews.com	83K	4.17%
irrawaddy.com	55K	2.77%
dvb.no	53K	2.68%
duwun.com.mm	38K	1.94%
thithtoolwin.com	30K	1.54%
xinhuamyanmar.com	26K	1.31%
dawnmanhon.com	23K	1.16%
blogspot.sg	21K	1.07%
hellosayarwon.com	21K	1.05%

Top 10 TLDs

Domain	Docs	% of total
com	1.3M	64.48%
org	173K	8.74%
com.mm	106K	5.33%
net	90K	4.52%
no	64K	3.21%
gov.mm	57K	2.88%
xyz	29K	1.46%
sg	21K	1.07%
media	13K	0.66%
co	11K	0.53%

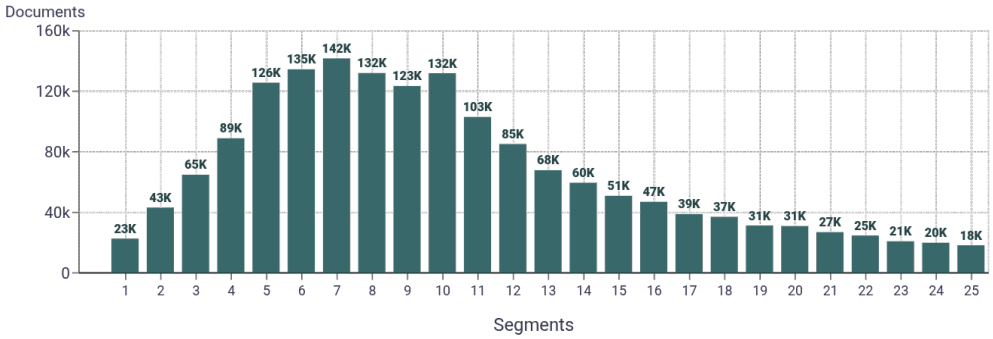
Register labels



MT:7.0% | 138K Documents

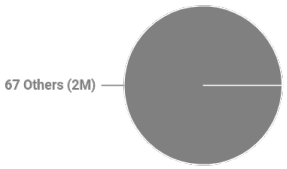
Documents size (in segments)

≤ 25 segments **84.45%** (1.7M documents)
> 25 segments **15.55%** (308K documents)



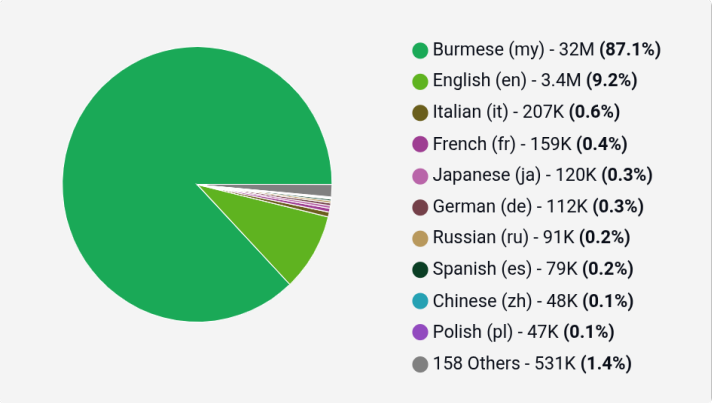
Document collections

CC = 87.28%
IA = 12.72%

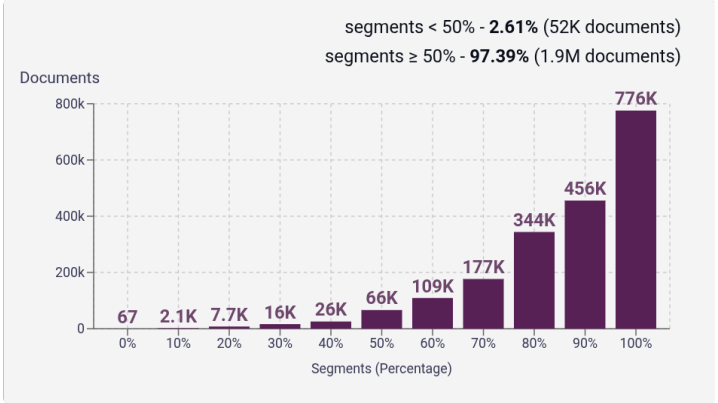


Language Distribution

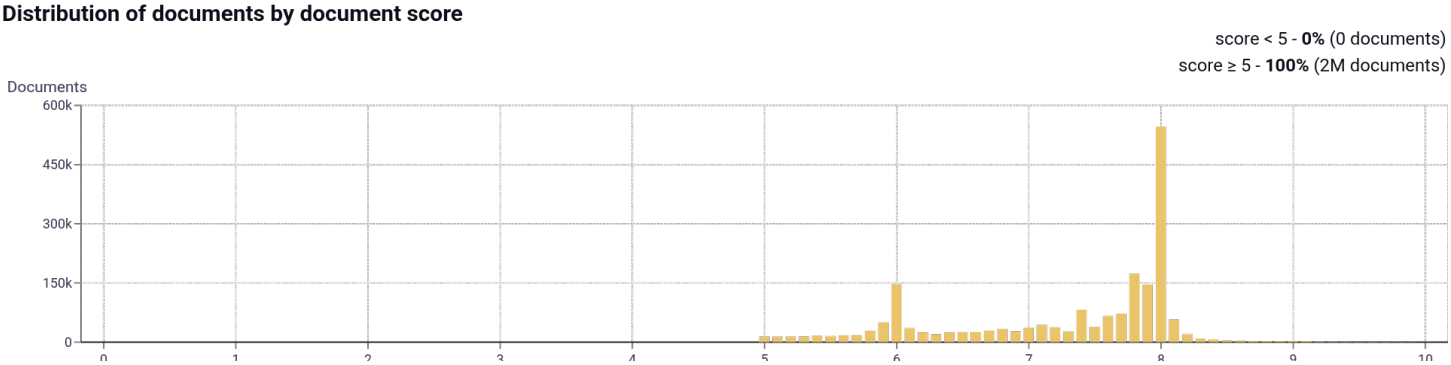
Number of segments in the Burmese (my) corpus



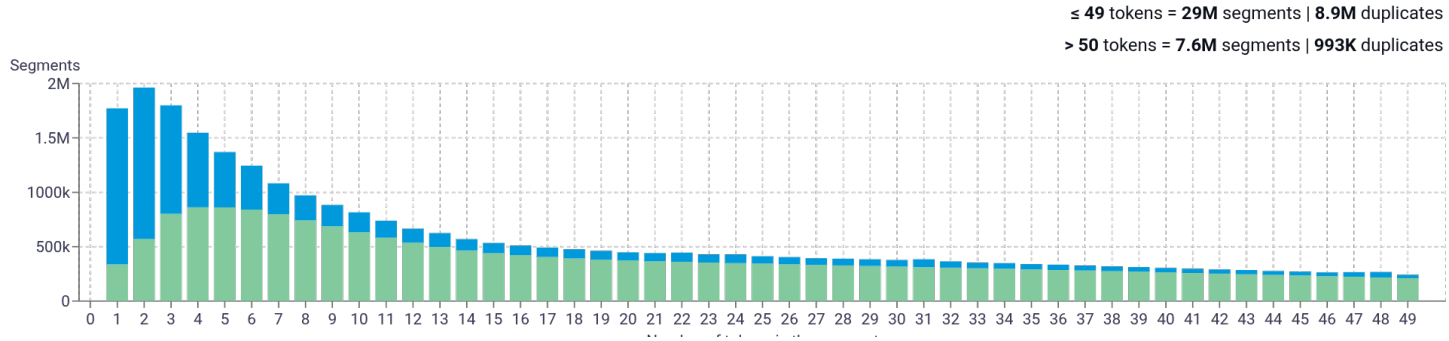
Percentage of segments in Burmese (my) inside documents



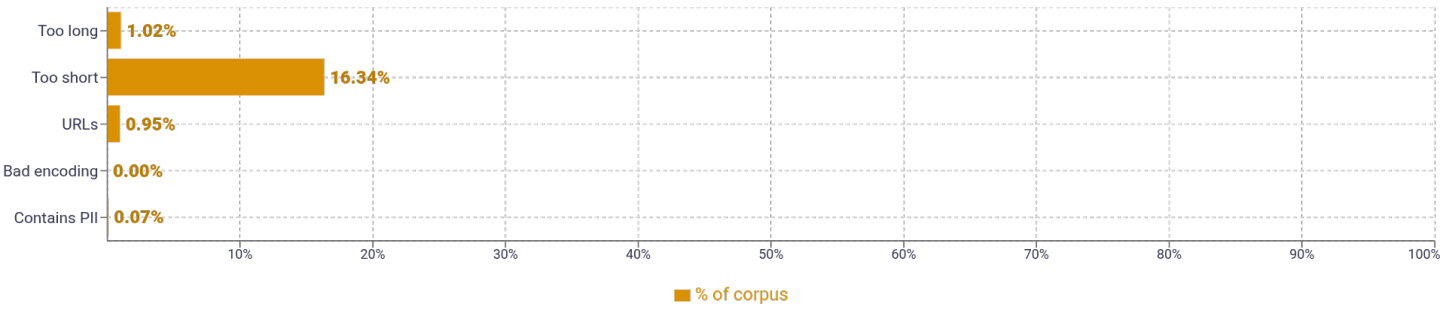
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ပါ | 26,005,591 · က | 25,127,287 · သည် | 16,145,726 · များ | 14,470,857 · ရ | 11,490,994
2	ပါ တယ် | 6,905,893 · ပါ တယ။ | 3,663,938 · ရ ပါ | 1,701,122 · တစ် ခု | 1,656,681 · ဖြစ် ပါ | 1,598,762
3	ဖြစ် ပါ တယ် | 1,270,571 · ရ ပါ တယ် | 636,912 · ပြော ပါ တယ် | 474,399 · နိုင် ပါ တယ် | 392,329 · ထား ပါ တယ် | 384,851
4	က ပြော ပါ တယ် | 290,049 · တာ ဖြစ် ပါ တယ် | 270,889 · မှာ ဖြစ် ပါ တယ် | 213,987 · ပဲ ဖြစ် ပါ တယ် | 183,013 · လို့သိ ရ ပါ တယ် | 173,506
5	တယ် လို့သိ ရ ပါ တယ် | 130,626 · တယ် လို့ ဆို ပါ တယ် | 87,480 · တာ ပဲ ဖြစ် ပါ တယ် | 63,728 · တွေ့ က ပြော ပါ တယ် | 53,055 · ရ မှာ ဖြစ် ပါ တယ် | 50,197

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				