HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tha_Thai.jsonl.tsv	9/29/2025	Thai (th)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
40,008,135	645,010,758	422,173,456 (65.45 %)	34.55%	27B	153,898,577,636	392.69 GB

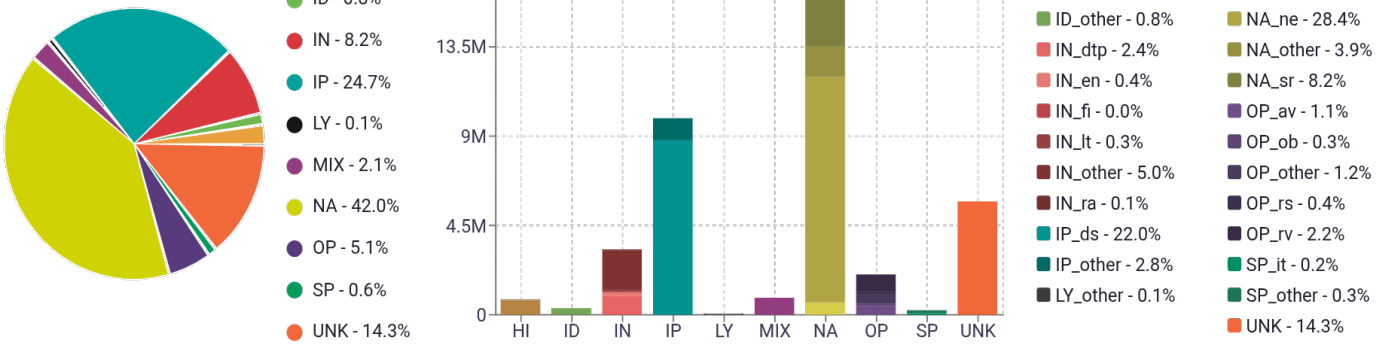
Top 10 domains

Domain	Docs	% of total
sanook.com	342K	0.85%
thairath.co.th	291K	0.73%
mthai.com	288K	0.72%
blogspot.com	228K	0.57%
ryt9.com	206K	0.52%
trueid.net	204K	0.51%
newswit.com	180K	0.45%
bangkokbiznews.com	175K	0.44%
wordpress.com	173K	0.43%
tripadvisor.com	172K	0.43%

Top 10 TLDs

Domain	Docs	% of total
com	24M	60.78%
net	2.7M	6.66%
xyz	2M	4.92%
co.th	1.9M	4.67%
org	1.7M	4.14%
co	645K	1.61%
go.th	591K	1.48%
ac.th	581K	1.45%
in.th	455K	1.14%
tw	449K	1.12%

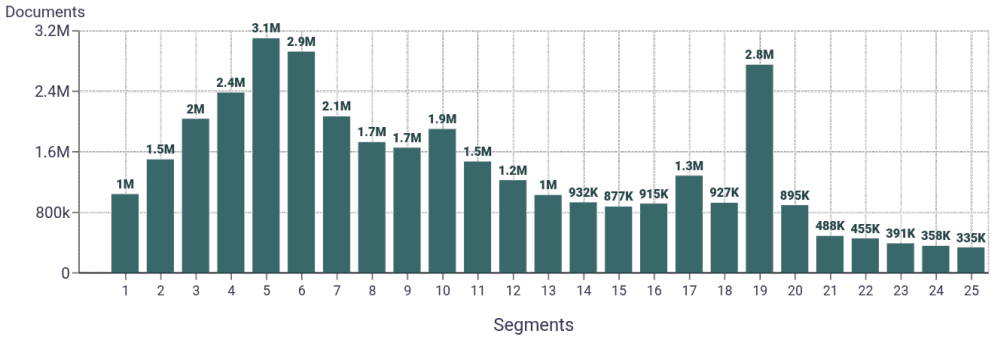
Register labels



MT:8.4% | 3.4M Documents

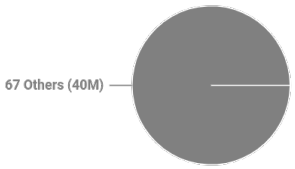
Documents size (in segments)

≤ 25 segments **86.67%** (35M documents)
> 25 segments **13.33%** (5.3M documents)



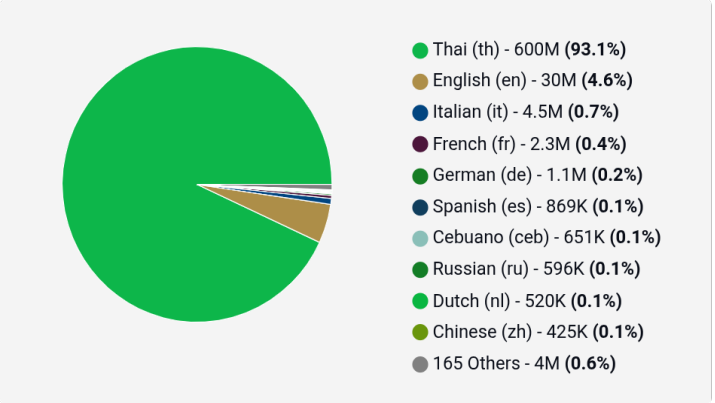
Document collections

CC = 91.71%
IA = 8.29%

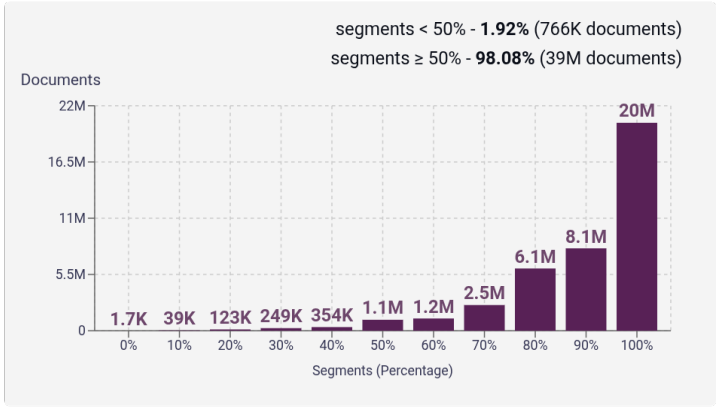


Language Distribution

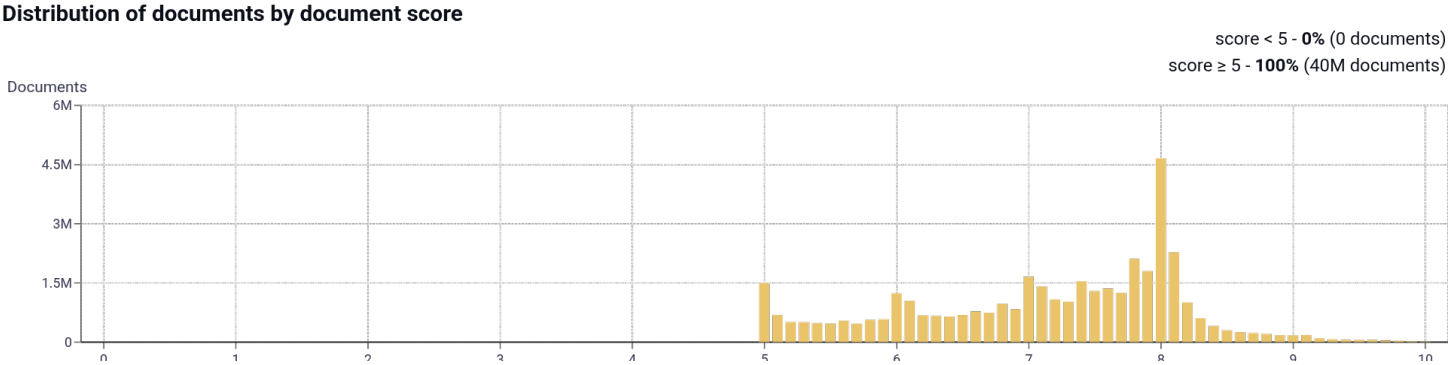
Number of segments in the Thai (th) corpus



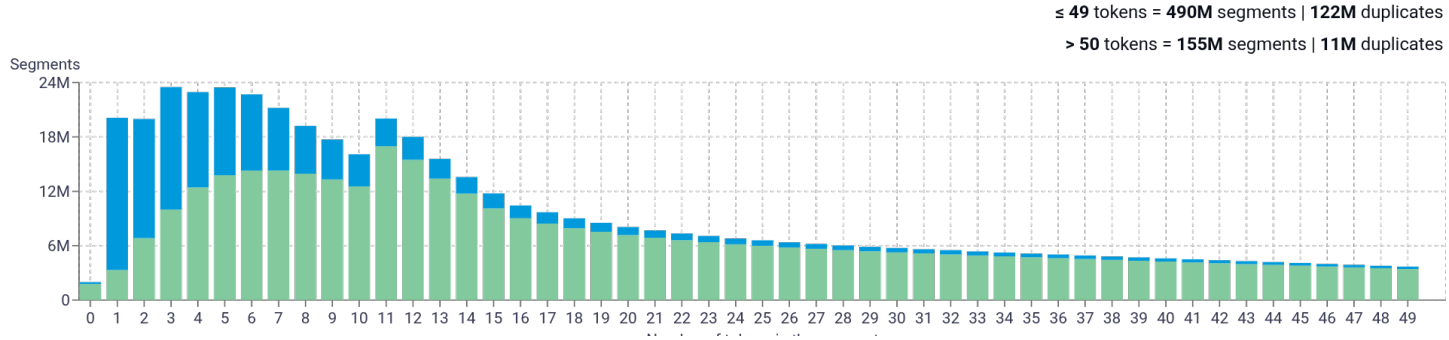
Percentage of segments in Thai (th) inside documents



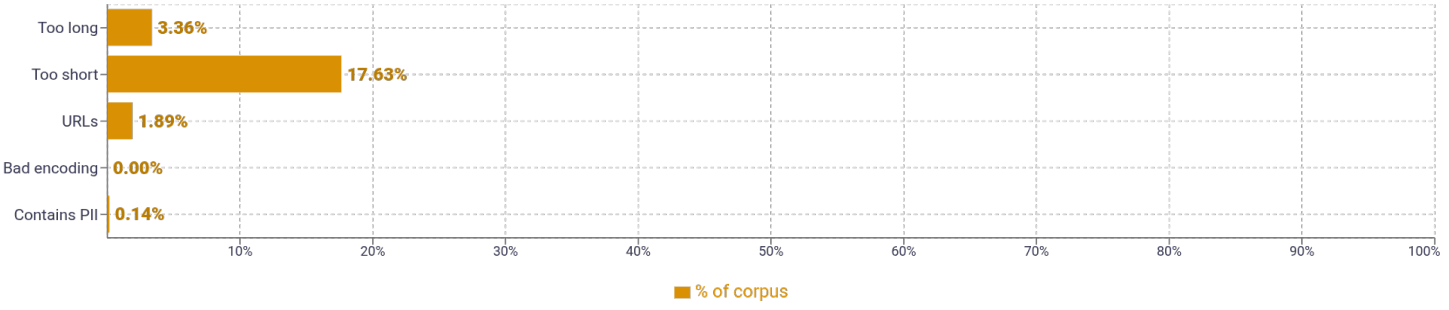
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	น | 102,113,081 · กา | 97,495,230 · ง | 94,305,902 · ก | 91,354,433 · ณ | 66,048,347
2	กา ร | 28,183,209 · email protected | 15,968,677 · online ads | 7,800,726 · tv ads | 7,800,585 · pg slot | 4,779,779
3	ก า ร | 1,371,478 · ก มา ย | 868,406 · ก ล่ า ว | 711,535 · ๐ ๐ ๐ | 685,169 · กร กฎา คม | 526,631
4	๐ ๐ ๐ ๐ | 669,248 · about the author naddanai | 256,543 · ก ำ ลั ง | 245,454 · true wallet ไม่มี ขั้น | 178,743 · wallet ไม่มี ขั้ น ต่ า | 135,403
5	๐ ๐ ๐ ๐ ๐ | 654,553 · ! ! ! ! ! | 104,558 · about the author naddanai อดีต | 102,791 · morning moon village สมัคร bitkub | 75,481 · author naddanai อดีตนักเขียนนิตยสาร inside united | 71,557

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				