HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lao_Laoo	9/17/2025	Lao

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
87,662	1,046,945	889,497 (84.96 %)	83M	287,295,602	750.04 MB

Top 10 domains

Domain	Docs	% of total
vnanet.vn	9.9K	11.28%
vovworld.vn	7.4K	8.39%
thoidai.com.vn	4.1K	4.68%
na.gov.la	3K	3.38%
laophattananews...	2.9K	3.31%
lnr.org.la	2.5K	2.90%
qdnd.vn	1.8K	2.02%
cri.cn	1.5K	1.68%
kongthap.gov.la	1.4K	1.62%
pasaxon.org.la	1.2K	1.33%

Top 10 TLDs

Domain	Docs	% of total
com	29K	33.01%
vn	19K	21.70%
gov.la	15K	16.87%
org.la	4.3K	4.88%
com.vn	4.1K	4.68%
org	3.6K	4.12%
com.la	2.2K	2.54%
net	1.8K	2.09%
cn	1.5K	1.73%
la	1.2K	1.42%

Register labels

- HI - 0.0%
- ID - 0.0%
- IN - 4.7%
- IP - 1.7%
- MIX - 0.3%
- NA - 67.3%
- OP - 0.5%
- SP - 0.3%
- UNK - 25.1%

- HI_other - 0.0%
- ID_other - 0.0%
- IN_dtp - 1.3%
- IN_en - 0.3%
- IN_lt - 0.4%
- IN_other - 2.7%
- IN_ra - 0.0%
- IP_ds - 1.2%
- IP_other - 0.5%
- MIX - 0.3%
- NA_nb - 0.1%
- NA_ne - 63.9%
- NA_other - 3.2%
- NA_sr - 0.1%
- OP_av - 0.0%
- OP_ob - 0.0%
- OP_other - 0.2%
- OP_rs - 0.3%
- OP_rv - 0.0%
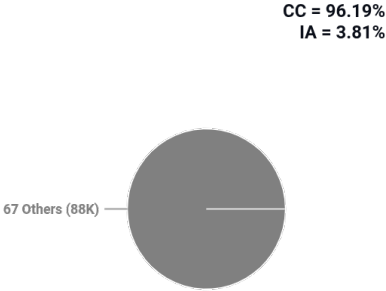- SP_it - 0.0%
- SP_other - 0.3%
- UNK - 25.1%
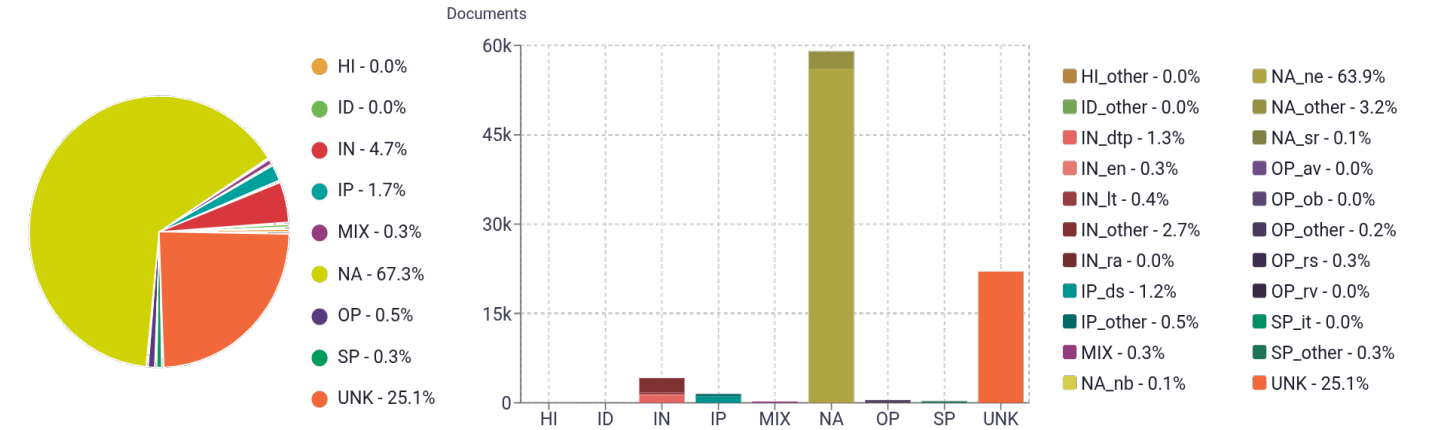
🤖 **MT**:23.0% | 20K Documents

Documents size (in segments) ⓘ
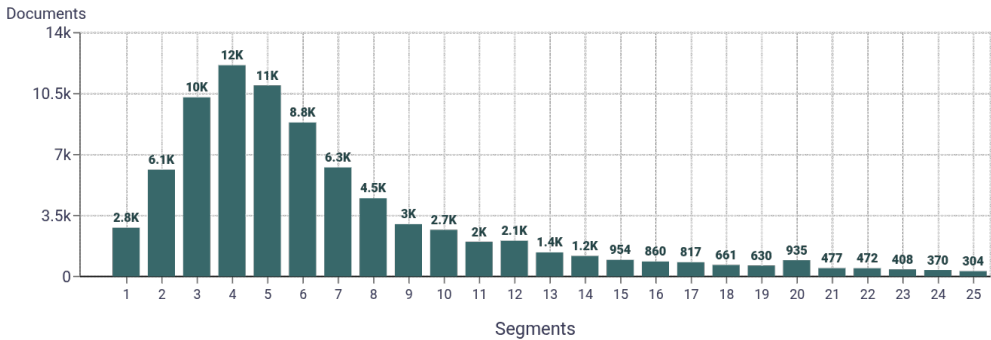
≤ 25 segments **92.5%** (81K documents)
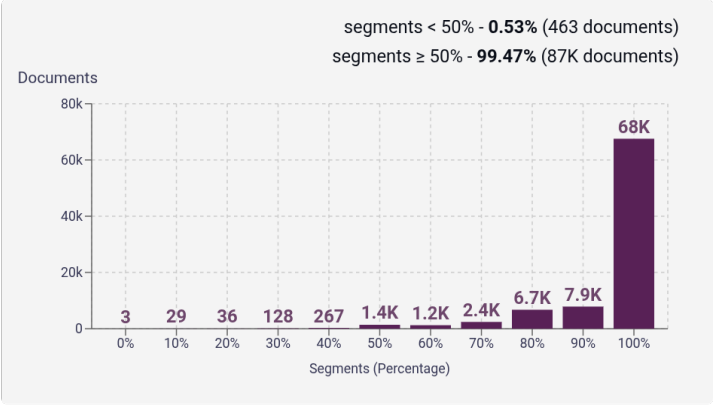**> 25** segments **7.5%** (6.6K documents)
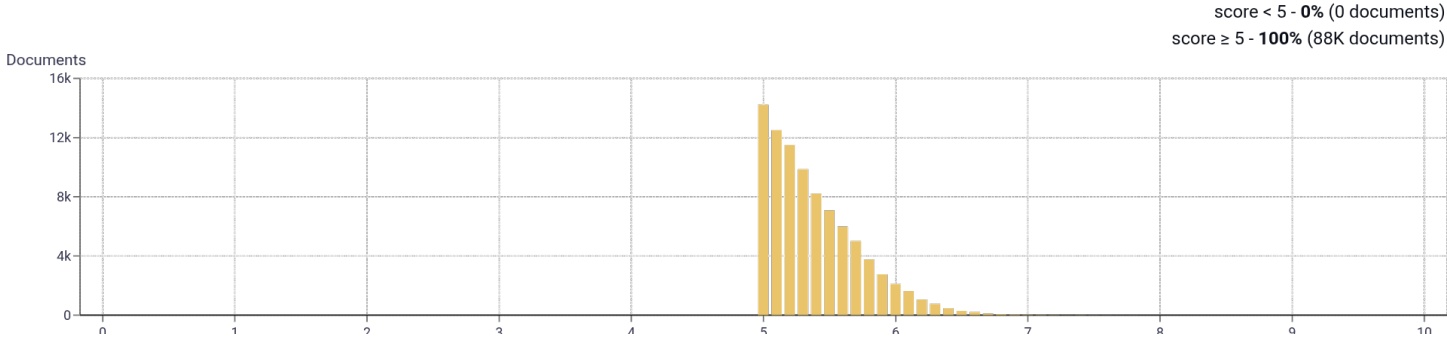

Document collections

CC = 96.19%
IA = 3.81%

67 Others (88K)

Language Distribution

Number of segments in the Lao corpus

- Lao - 964K **(92.1%)**
- English - 25K **(2.3%)**
- Ukrainian - 10K **(1.0%)**
- Italian - 9.2K **(0.9%)**
- Greek - 7.3K **(0.7%)**
- French - 2.7K **(0.3%)**
- German - 2.6K **(0.2%)**
- Slovenian - 2.1K **(0.2%)**
- Kannada - 2K **(0.2%)**
- Tosk Albanian - 2K **(0.2%)**
- 138 Others - 20K **(1.9%)**
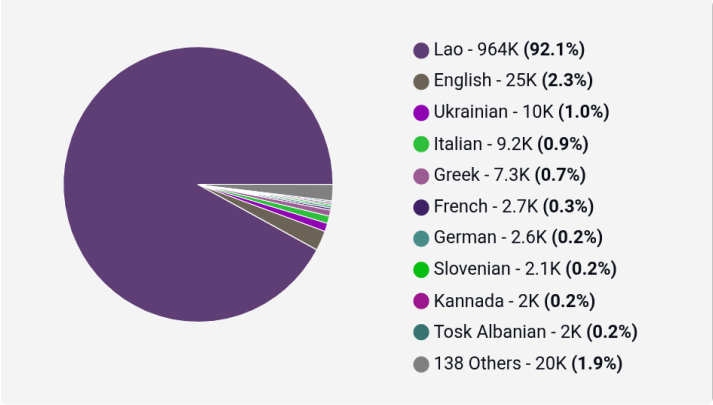
Percentage of segments in Lao inside documents

segments < 50% - **0.53%** (463 documents)
segments ≥ 50% - **99.47%** (87K documents)

Documents

0%	3
10%	29
20%	36
30%	128
40%	267
50%	1.4K
60%	1.2K
70%	2.4K
80%	6.7K
90%	7.9K
100%	68K

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (88K documents)

Documents

Segment length distribution by token

≤ 49 tokens = 532K segments | 124K duplicates
> 50 tokens = 515K segments | 34K duplicates

Segments

Segment noise distribution

Too long	**2.94%**
Too short	**17.79%**
URLs	**0.98%**
Bad encoding	**0.01%**
Contains PII	**0.17%**

■ % of corpus
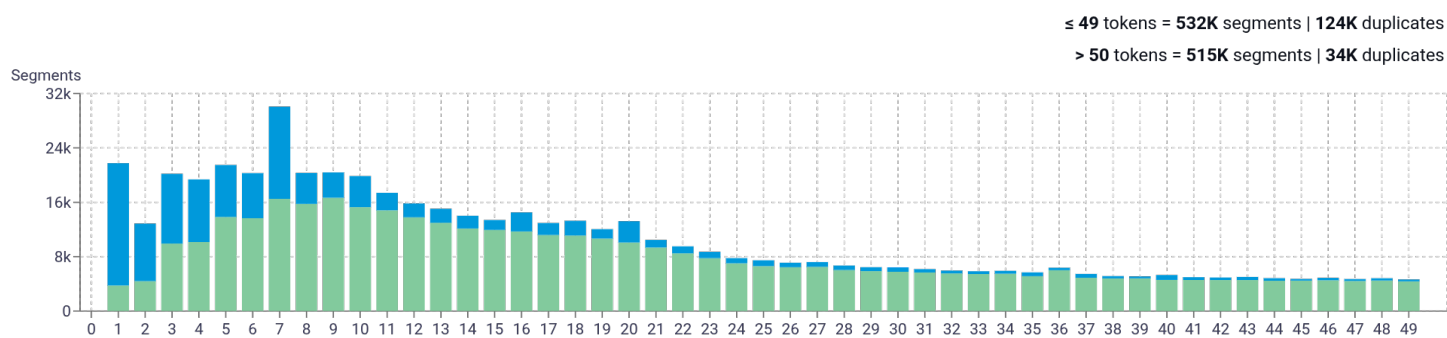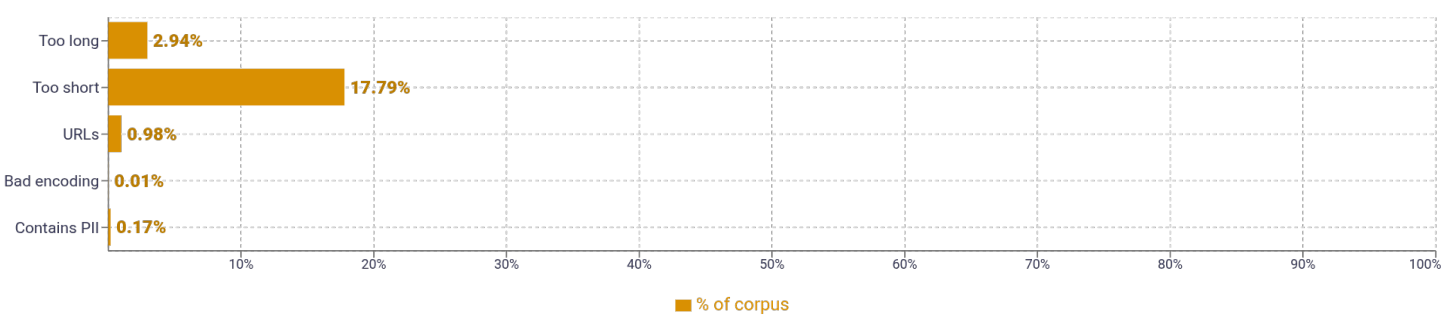
Frequent n-grams

SIZE	N-GRAMS	
1	ໆ | 670,781 ທີ່ | 529,686 ອ | 423,478 ສໍ | 244,395 ວາ | 198,784	
2	ສໍາ | 182,218 ກອງ ປະຊຸມ | 138,578 ບໍ່ໆ | 101,173 ໆ ລັບ | 94,286 ຄໍາ ໆ | 62,578	
3	ສໍາ ລັບ | 94,239 ສໍາ ຄັນ | 39,165 ບໍ່ໆ ໃຊ້ | 37,384 ສໍາ ເລັດ | 23,694 ຈໍາ ນວນ | 23,172	
4	ຄະນະ ບໍລິຫານ ງານ ສູນກາງ | 8,693 ທໍ່ ມະ ຊາດ | 8,651 ແບວ ລາວ ສ້າງ ຊາດ | 8,579 ກອງ ປະຊຸມ ສຸດ ຍອດ | 7,307 ກົດ ຈະ ກໍາ | 7,213	
5	ຄະນະ ບໍລິຫານ ງານ ສູນກາງ ພັກ | 7,050 ຄໍາ ແນະ ບໍ່ໆ | 4,498 ຄະນະ ໂຄສະນາ ອົບຮົມ ສູນກາງ ພັກ | 4,182 ສູນກາງ ແບວ ລາວ ສ້າງ ຊາດ | 3,818 ກອງ ປະຊຸມ ໃຫຍ່ ຄັ້ງ ທີ | 3,398	

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				