

General overview

Corpus	Date	Language
hplt-v3-lao_Lao	9/17/2025	Lao (lo)

Volumes

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

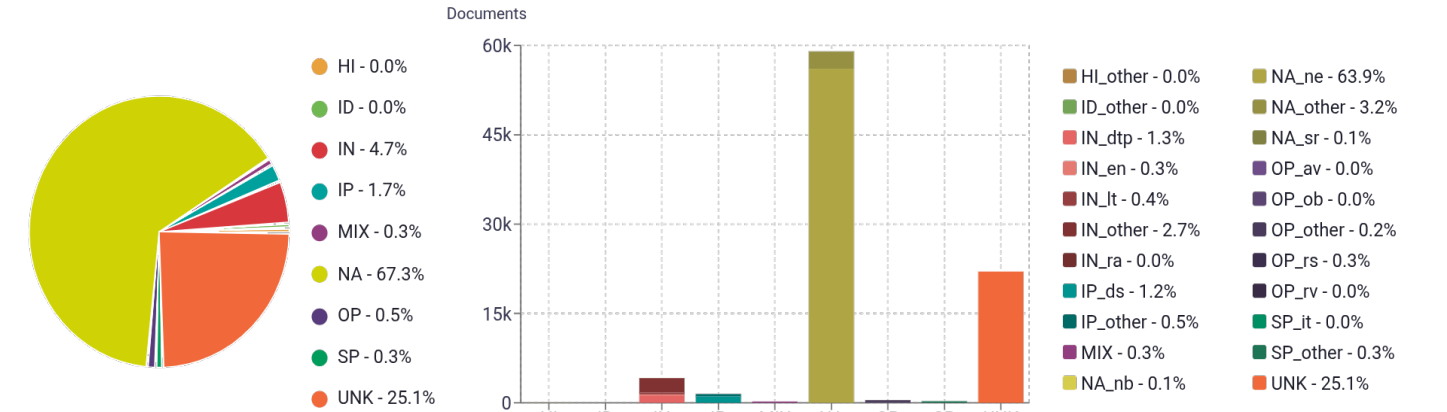
Top 10 domains

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

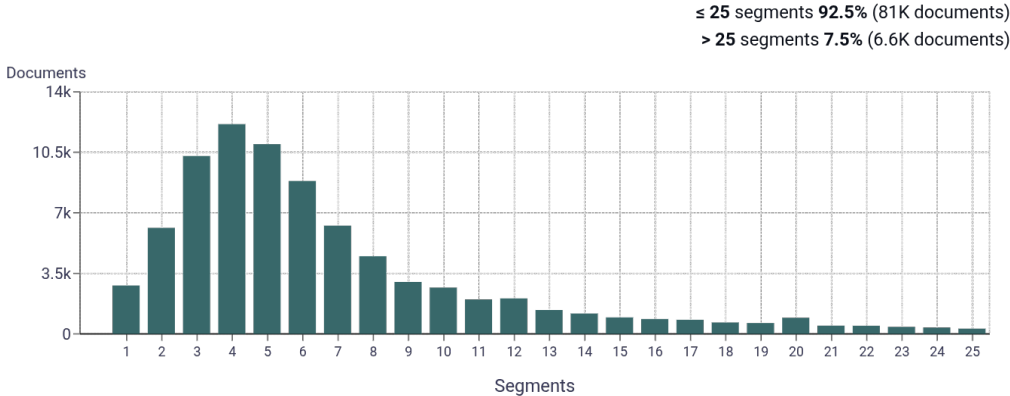
Domain	Docs	% of total
com	29K	33.01%
vn	19K	21.70%
gov.la	15K	16.87%
org.la	4.3K	4.88%
com.vn	4.1K	4.68%
org	3.6K	4.12%
com.la	2.2K	2.54%
net	1.8K	2.09%
cn	1.5K	1.73%
la	1.2K	1.42%

Register labels

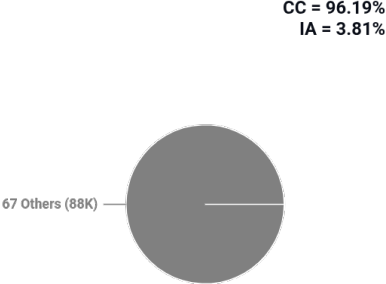


MT:23.0% | 20K Documents

Documents size (in segments)

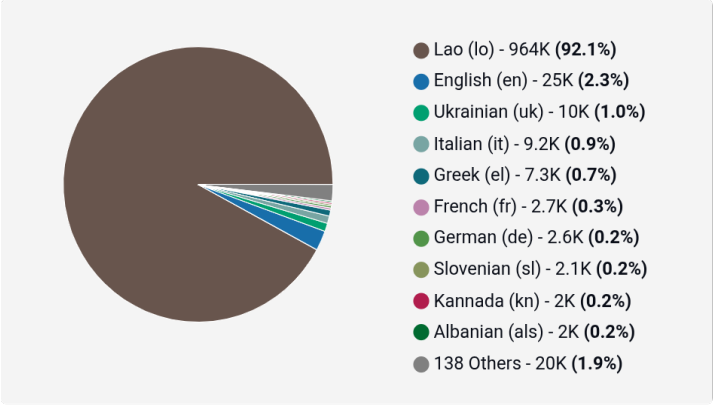


Document collections

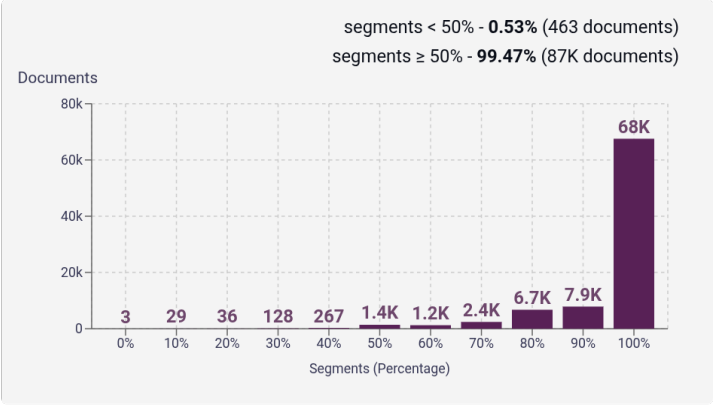


Language Distribution

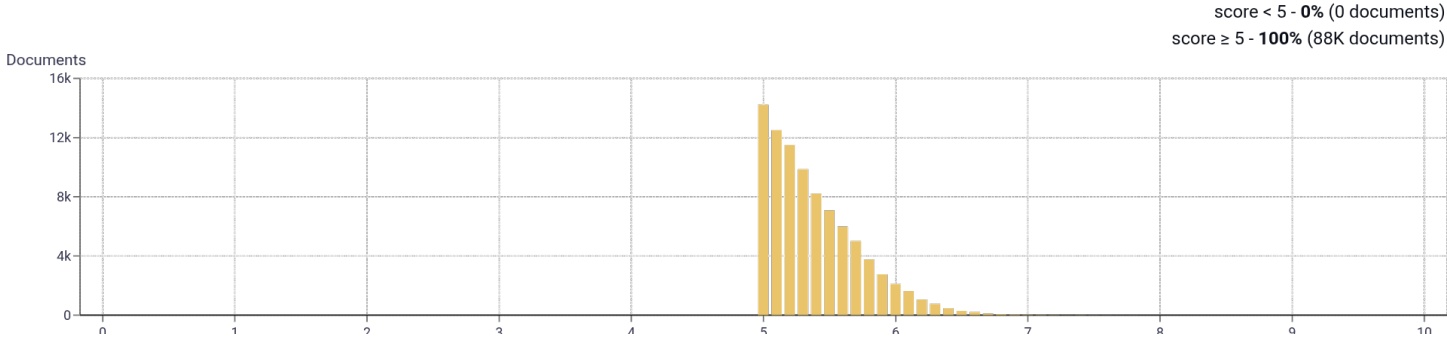
Number of segments in the Lao (lo) corpus



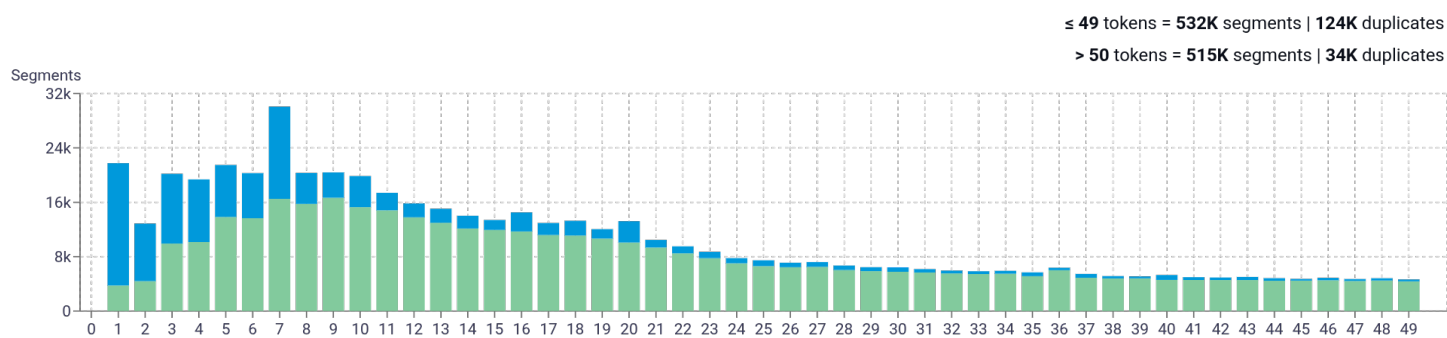
Percentage of segments in Lao (lo) inside documents



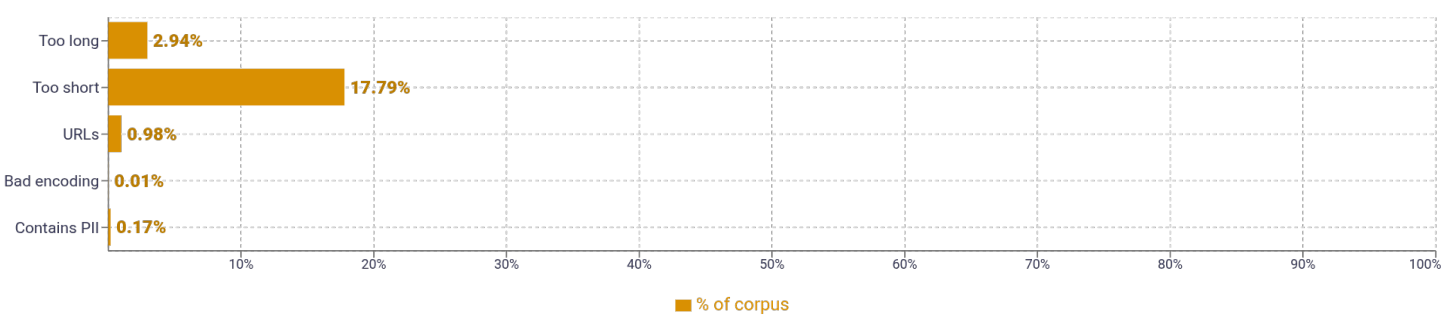
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	າ   670,781    ຕີ   529,686    ອ   423,478    ສໍ   244,395    ວາ   198,784	
2	ສໍາ   182,218    ກອງ ປະຊຸມ   138,578    ນໍາ   101,173    າລັບ   94,286    ຄໍາ   62,578	
3	ສໍາລັບ   94,239    ສໍາຄັນ   39,165    ນໍາໃຊ້   37,384    ສໍາເລັດ   23,694    ຈໍານວນ   23,172	
4	ຄະນະ ບໍລິຫານ ງານ ສູນກາງ   8,693    ທໍາມະ ຊາດ   8,651    ແນວ ລາວ ສ້າງ ຊາດ   8,579    ກອງ ປະຊຸມ ສຸດ ຍອດ   7,307    ກິດ ຈະ ກໍາ   7,213	
5	ຄະນະ ບໍລິຫານ ງານ ສູນກາງ ພັກ   7,050    ຄໍາ ແນະ ນໍາ   4,498    ຄະນະ ໂຄສະນາ ອົບຮົມ ສູນກາງ ພັກ   4,182    ສູນກາງ ແນວ ລາວ ສ້າງ ຊາດ   3,818 ກອງ ປະຊຸມ ໃຫຍ່ ຄັ້ງ ທີ່   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 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>, <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				