HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
lao_Laoo.jsonl.tsv	9/23/2024	Lao (lo)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
29,504	319,953	221,388 (69.19 %)	22M	84,390,143	221.5 MB

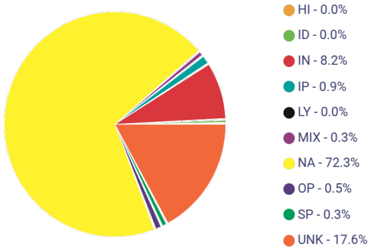
Top 10 domains

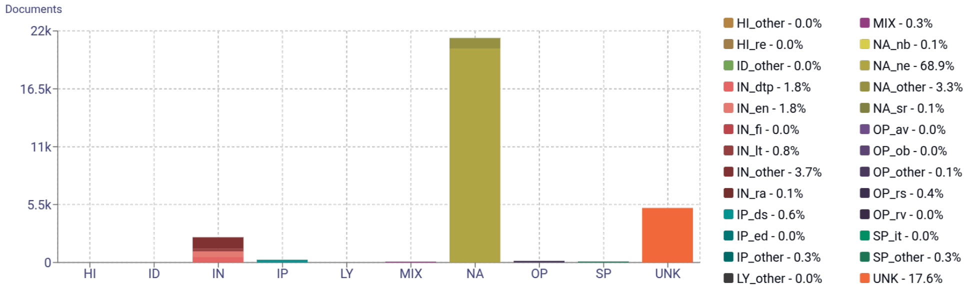
Domain	Docs	% of total
vnanet.vn	5.2K	17.55%
vovworld.vn	2K	6.66%
na.gov.la	1.3K	4.29%
lnr.org.la	1.1K	3.77%
lsr.com.la	981	3.32%
thoidai.com.vn	941	3.19%
wikipedia.org	709	2.40%
tapchicongsan.o...	534	1.81%
cri.cn	473	1.60%
nuol.edu.la	358	1.21%

Top 10 TLDs

Domain	Docs	% of total
vn	7.5K	25.38%
com	7.4K	25.08%
gov.la	5.6K	19.11%
org	2K	6.67%
org.la	1.5K	5.04%
com.la	1.4K	4.66%
com.vn	941	3.19%
edu.la	620	2.10%
org.vn	534	1.81%
cn	487	1.65%

Register labels

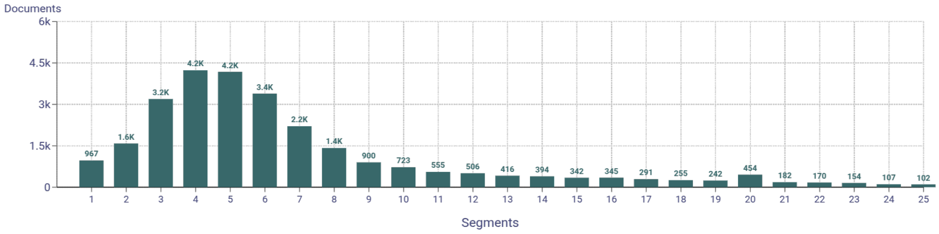




🤖 **MT**:14.7% | 4.3K Documents

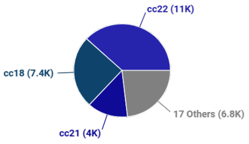
Documents size (in segments)

<= 25 segments **92.57%** (27K documents)
> 25 segments **7.43%** (2.2K documents)



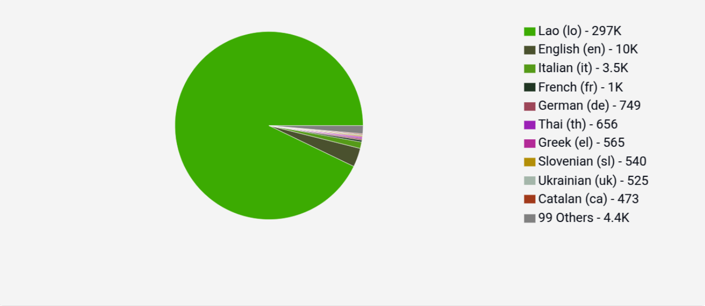
Documents by collection

CC = 83.06%
IA = 16.94%

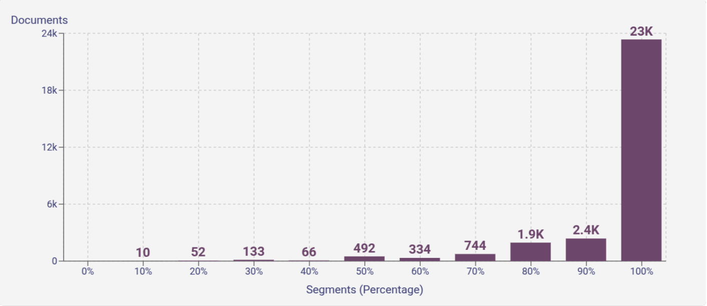


Language Distribution

Number of segments in the Lao (lo) corpus

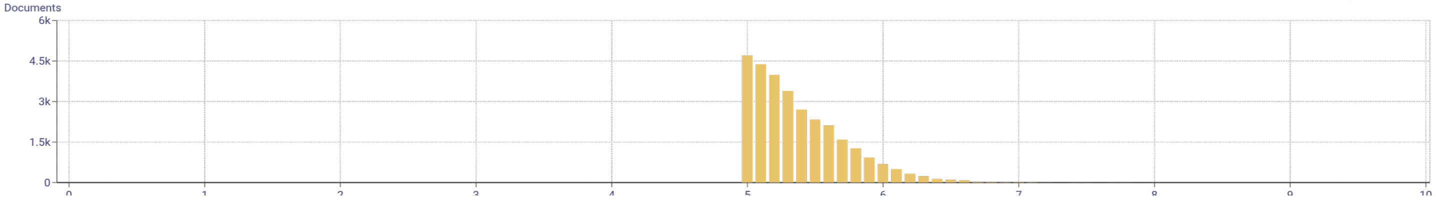


Percentage of segments in Lao (lo) inside documents



Distribution of documents by document score

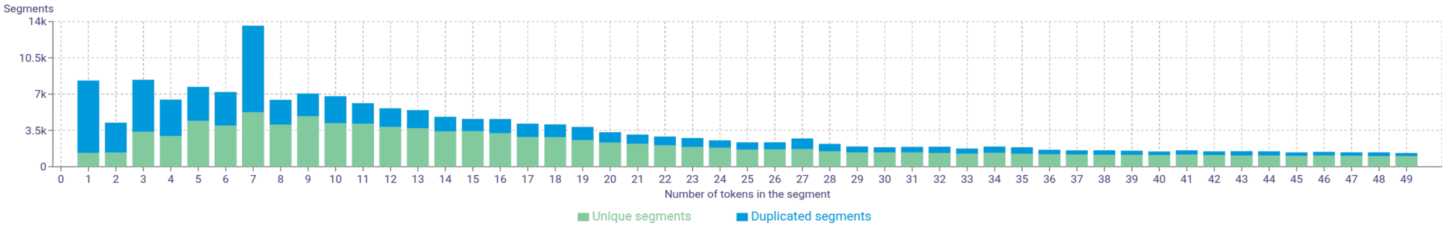
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (30K documents)



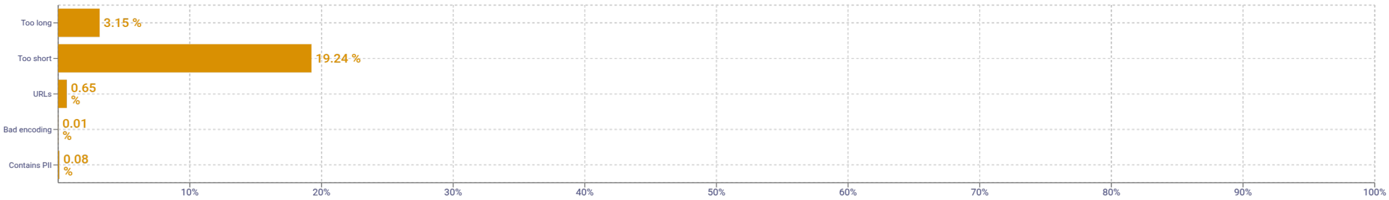
Segment length distribution by token

≤ **49** tokens = **107K** segments | **70K** duplicates

> **50** tokens = **142K** segments | **28K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ໆ | 156290 ທ່ານ | 129958 ລາວ | 92003 ແມ່ນ | 84606 ພັກ | 74319
2	ກອງ ປະຊຸມ | 53088 ສໍ ໆ | 42316 ສະພາ ແຫ່ງຊາດ | 25340 ນໍ ໆ | 24536 ໆ ລັບ | 20624
3	ສໍ ໆ ລັບ | 20609 ສໍ ໆ ຄັນ | 8441 ນໍ ໆ ໃຊ້ | 8161 ຄະນະ ບໍລິຫານ ງານ | 7607 ຈໍ ໆ ນອນ | 7230
4	ຫງ ວຽນ ຊວນ ຟຸກ | 3904 ຄະນະ ບໍລິຫານ ງານ ສູນກາງ | 3721 ແນວ ລາວ ສ້າງ ຊາດ | 3676 ບໍລິຫານ ງານ ສູນກາງ ພັກ | 3156 ກົມ ການເມືອງ ສູນກາງ ພັກ | 2971
5	ຄະນະ ບໍລິຫານ ງານ ສູນກາງ ພັກ | 3083 ກອງ ປະຊຸມ ໃຫຍ່ ຄັ້ງ ທີ | 1614 ສູນກາງ ແນວ ລາວ ສ້າງ ຊາດ | 1590 ຄະນະ ໂຄສະນາ ອົບຮົມ ສູນກາງ ພັກ | 1588 ກຳມະການ ກົມ ການເມືອງ ສູນກາງ ພັກ | 1328

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av