HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tgl_Latn.jsonl.tsv	9/6/2024	Filipino (tl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,868,959	52,879,871	27,036,754 (51.13 %)	1.6B	8,079,611,643	7.57 GB

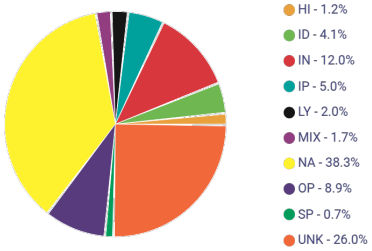
Top 10 domains

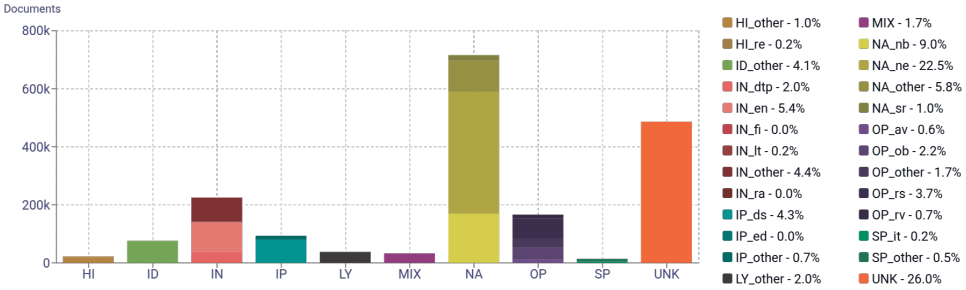
Domain	Docs	% of total
blogspot.com	119K	6.38%
wikipedia.org	94K	5.04%
remate.ph	77K	4.13%
wordpress.com	51K	2.70%
pinoyparazzi.com	35K	1.88%
pep.ph	30K	1.61%
dwiz882am.com	27K	1.46%
abante.com.ph	25K	1.32%
abs-cbn.com	23K	1.25%
hataw tabloid.com	19K	1.02%

Top 10 TLDs

Domain	Docs	% of total
com	968K	51.81%
org	197K	10.53%
ph	163K	8.73%
net	91K	4.89%
ru	82K	4.37%
com.ph	78K	4.15%
pl	21K	1.13%
gov.ph	21K	1.10%
info	19K	1.04%
tk	18K	0.97%

Register labels

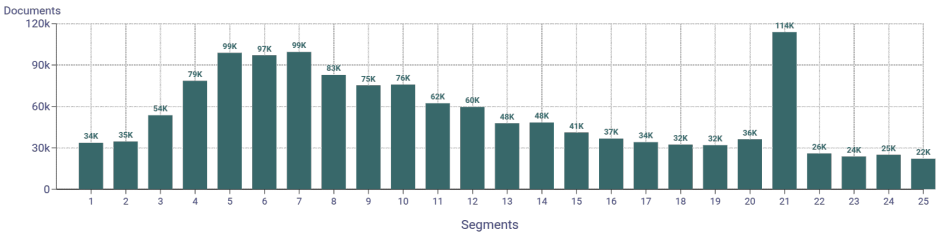




MT:23.4% | 437K Documents

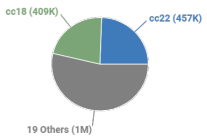
Documents size (in segments)

<= 25 segments **73.33%** (1.4M documents)
> 25 segments **26.67%** (499K documents)



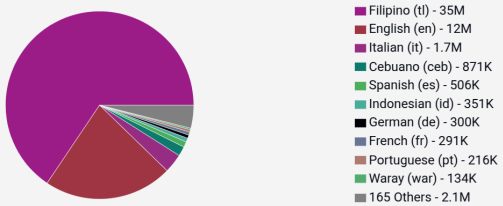
Documents by collection

CC = 71.02%
IA = 28.98%

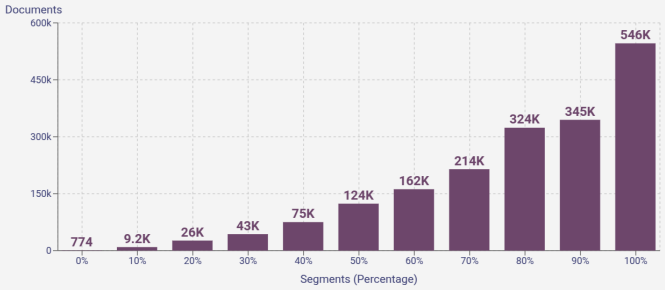


Language Distribution

Number of segments in the Filipino (tl) corpus

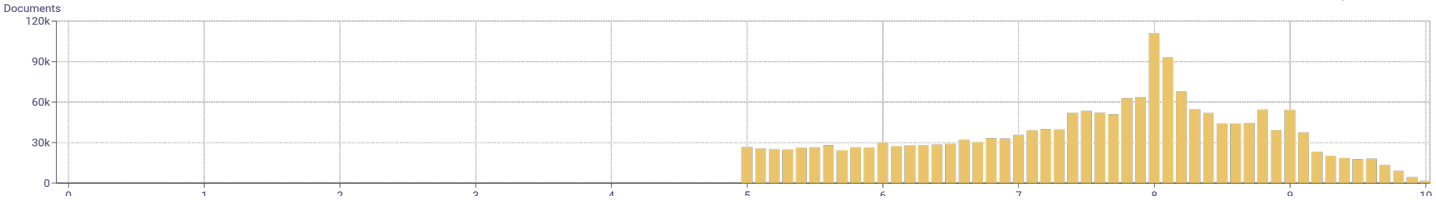


Percentage of segments in Filipino (tl) inside documents



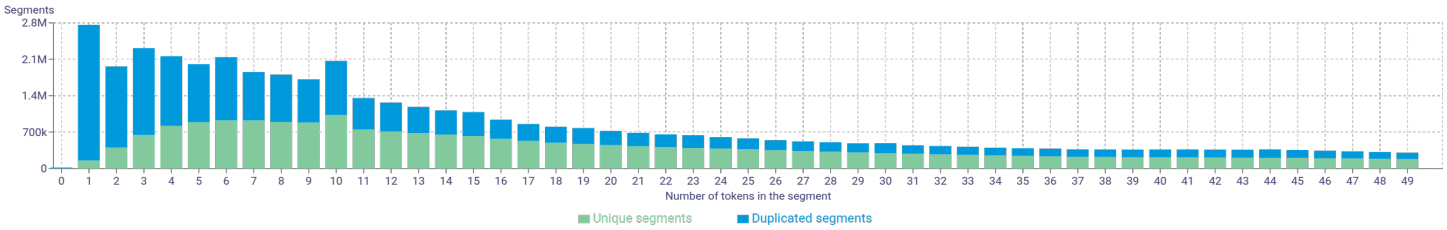
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.9M documents)

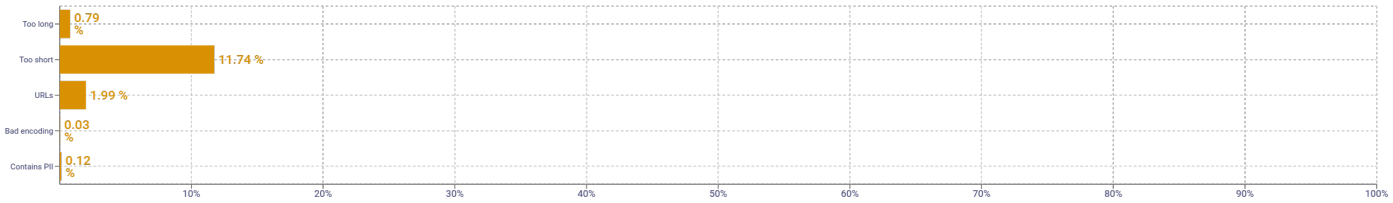


Segment length distribution by token

≤ 49 tokens = **21M** segments | **22M** duplicates
> 50 tokens = **9.3M** segments | **3.6M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	si | 5734355 · mo | 5100988 · the | 4926297 · lang | 4737829 · naman | 4033321
2	of the | 635409 · kuko halamang-singaw | 386309 · in the | 383085 · t ibang | 293785 · 't ibang | 251634
3	pagbaba ng timbang | 221070 · pati na rin | 174092 · mawalan ng timbang | 150615 · you must be | 139435 · logged in to | 138324
4	you must be logged | 138211 · must be logged in | 138210 · be logged in to | 138209 · to post a comment | 137377 · in to post a | 137336
5	you must be logged in | 138209 · must be logged in to | 138200 · in to post a comment | 137326 · logged in to post a | 137311 · be logged in to post | 137308

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