HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-pag_Latn	9/18/2025	Pangasinan (pag)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,496	171,548	155,848 (90.85 %)	8.9M	41,871,723	40.42 MB

Top 10 domains

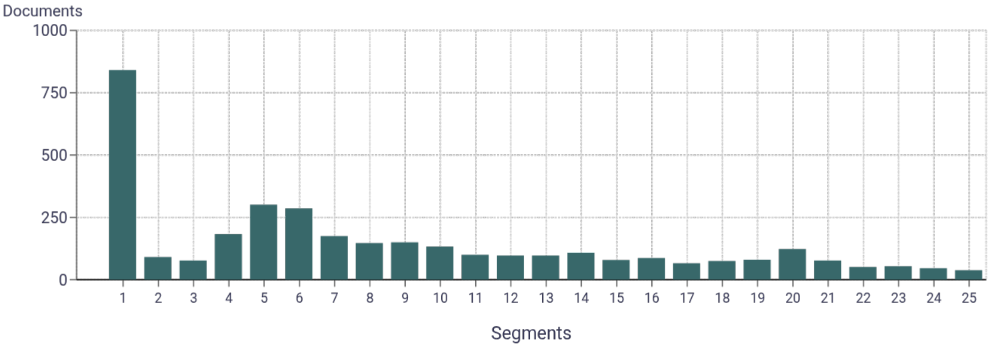
Domain	Docs	% of total
jw.org	1.9K	41.81%
bible.is	829	18.44%
bomboradyo.com	316	7.03%
pacificbibles.org	260	5.78%
ebible.org	249	5.54%
wikipedia.org	203	4.52%
blogspot.com	84	1.87%
bibles.org	84	1.87%
webonary.org	60	1.33%
wordpress.com	56	1.25%

Top 10 TLDs

Domain	Docs	% of total
org	2.8K	61.99%
is	829	18.44%
com	685	15.24%
gov.ph	64	1.42%
pk	30	0.67%
net	15	0.33%
info	15	0.33%
jp	11	0.24%
center	11	0.24%
tw	9	0.20%

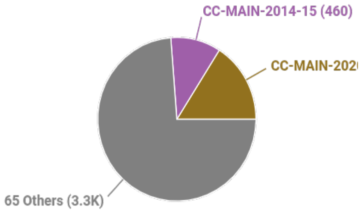
Documents size (in segments) ⓘ

≤ **25** segments **79.23%** (3.6K documents)
> 25 segments **20.77%** (934 documents)



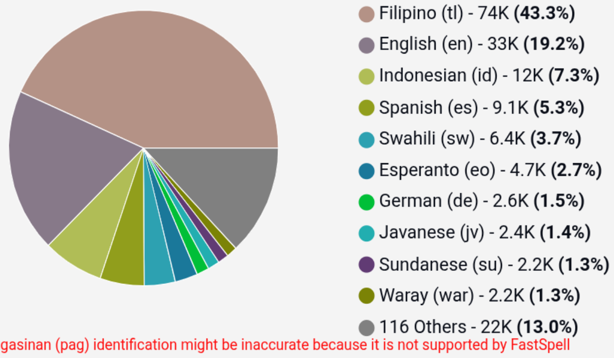
Document collections

CC = 83.39%
IA = 16.61%



Language Distribution

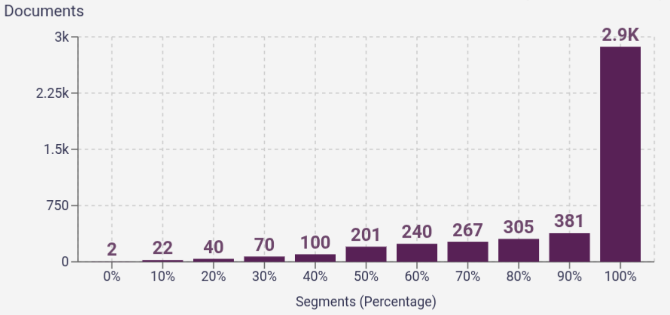
Number of segments in the Pangasinan (pag) corpus



*Pangasinan (pag) identification might be inaccurate because it is not supported by FastSpell

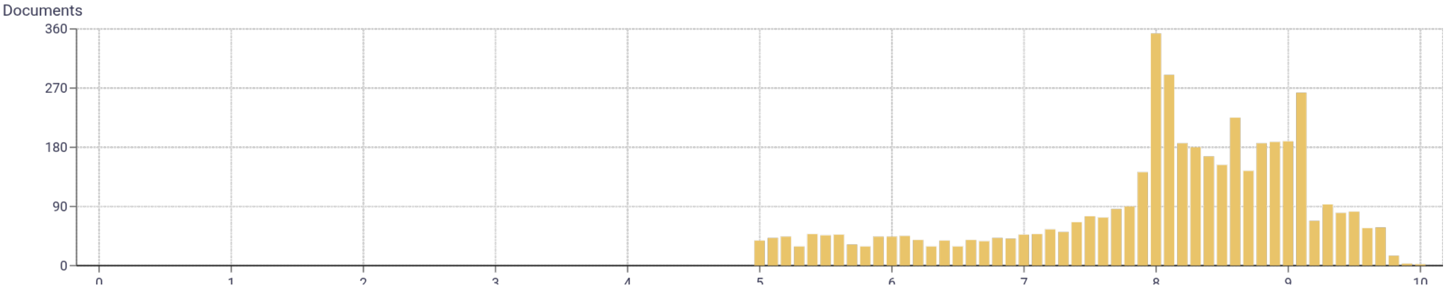
Percentage of segments in Pangasinan (pag) inside documents

segments < 50% - **5.20%** (234 documents)
segments ≥ 50% - **94.80%** (4.3K documents)



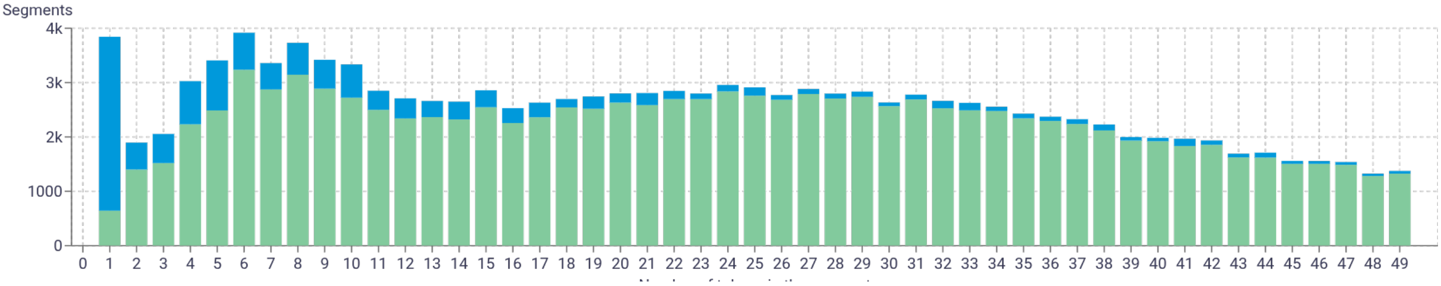
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4.5K documents)

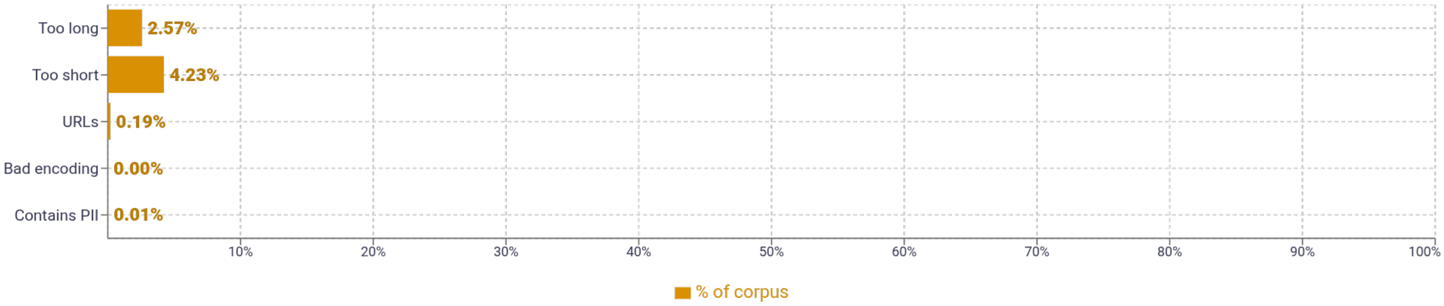


Segment length distribution by token

≤ **49** tokens = **127K** segments | **14K** duplicates
> 50 tokens = **44K** segments | **1.4K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ya | 187,626 so | 158,376 ko | 108,323 nin | 92,142 ed | 89,462
2	so manga | 15,592 ko manga | 10,138 ed saray | 9,103 sii ko | 8,385 si jesus | 8,364
3	nen apo shiyos | 3,127 ed si jehova | 2,169 ago so manga | 2,075 ta pigaw nin | 1,890 hi apo hisos | 1,547
4	ja kowan to ey | 732 aya pen o ba | 628 saray tasi nen jehova | 620 manga oripn o allāh | 561 nen apo shiyos ja | 550
5	nen apo shiyos son si | 268 allah a lebi a mamaapaar | 248 ispirito santo nen apo shiyos | 231 lebi a maporo a tohan | 219 pinili ni apo namalyari hên | 192

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